

Proceedings of the EAA Spatial Audio Signal Processing symposium

Markus Noisternig, Brian F. G. Katz, Boaz Rafaely

▶ **To cite this version:**

Markus Noisternig, Brian F. G. Katz, Boaz Rafaely (Dir.). Proceedings of the EAA Spatial Audio Signal Processing symposium. 2019, ⟨10.25836/sasp.2019.99⟩. ⟨hal-02275032⟩

HAL Id: hal-02275032

https://hal.science/hal-02275032v1

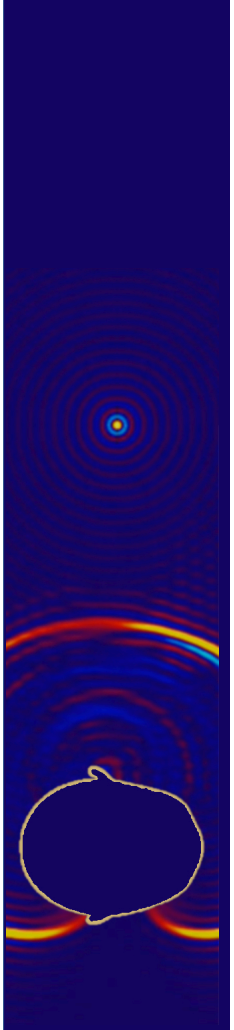
Submitted on 2 Sep 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



HAL Authorization



EAA SASP 2019

PROCEEDINGS OF THE EAA SPATIAL AUDIO SIGNAL PROCESSING SYMPOSIUM

SEPTEMBER 6-7, 2019

SORBONNE UNIVERSITÉ, PARIS, FRANCE

EAA · SFA · CNRS · ircam Centre Pompidou · LAM Institut Jean le Rond d'Alembert · SORBONNE UNIVERSITÉ

PROCEEDINGS OF THE EAA SPATIAL AUDIO SIGNAL PROCESSING SYMPOSIUM

SEPTEMBER 6-7, 2019

SORBONNE UNIVERSITÉ, PARIS, FRANCE

Published by

Sorbonne Université

https://www.sorbonne-universite.fr

DOI: 10.25836/sasp.2019.99

Copyright

Proceedings of the EAA Spatial Audio Signal Processing Symposium

Editors: Markus Noisternig, Brian FG Katz, Boaz Rafaely

© 2019, All copyrights remain with the authors.

Credits

Proceedings edited using DC/ConfOrg (http://conforg.fr).

EAA SASP 2019 is supported by:

	European Acoustics Association	https://euracoustics.org
	Société Française d'Acoustique	https://sfa.asso.fr
	Centre National de la Recherche Scientifique	https://www.cnrs.fr
	IRCAM – Centre Pompidou	https://www.ircam.fr
	Institut Jean le Rond d'Alembert	http://www.dalembert.upmc.fr/lam
	Sorbonne Université	https://www.sorbonne-universite.fr

Organizing Committees

Conference Co-Chairs

Markus Noisternig	STMS IRCAM, CNRS, Sorbonne Université
Brian FG Katz	Institut Jean Le Rond d'Alembert, Sorbonne Université, CNRS
Boaz Rafaely	Ben Gurion University of the Negev

Publication Chair

Olivier Warusfel	STMS IRCAM, CNRS, Sorbonne Université

Technical Chair

Thibaut Carpentier	STMS IRCAM, CNRS, Sorbonne Université

Financial Chair

Jean-Michel Ville	Université de Technologie de Compiègne

Administration Co-Chairs

Sandrine Bandeira	Institut Jean Le Rond d'Alembert, Sorbonne Université
Sylvie Benoit	STMS IRCAM, CNRS, Sorbonne Université

Local Organizing Committee

Fatiha Benrezzak	Coordination SFA
Didier Cassereau	Conference Management System
Evelyne Dewayse	Coordination SFA
Claire Marquet	Communication
Markus Noisternig	Website

TABLE OF CONTENTS

Session R009 - Spatial audio signal processing

Deep-sound field analysis for upscaling ambisonic signals
G. Routray, S. Basu, P. Baldev and R.M. Hegde 1

First-order ambisonic coding with quaternion-based interpolation of PCA rotation matrices
P. Mahé, S. Ragot and S. Marchand 7

Block-sparse approach for the identification of complex sound sources in a room
H. Demontis, F. Ollivier and J. Marchal 13

Online DOA estimation using real eigenbeam ESPRIT with propagation vector matching
A. Herzog and E. Habets 19

Beamforming with double-sided, acoustically hard planar arrays
S. Berge 25

Session R011 - Modeling, simulation, and analysis

Virtual acoustic rendering by state wave synthesis
E. Maestre, G.P. Scavone and J.O. Smith 31

Simulative investigation of required spatial source resolution in directional room impulse response measurements
J. Klein and M. Vorländer 37

Assessing the anisotropic features of spatial impulse responses
B. Alary, P. Massé, V. Välimäki and M. Noisternig 43

Session R012 - HRTF analysis

How wearing headgear affects measured head-related transfer functions
C. Pörschmann, J.M. Arend and R. Gillioz 49

Influence of vision on short-term sound localization training with non-individualized HRTF
T. Bouchara, T.-G. Bara, P.-L. Weiss and A. Guilbert 55

The influence of different BRIR modification techniques on externalization and sound quality
P.M. Giller, F. Wendt and R. Höldrich 61

Session R013 - Perceptual evaluation of spatial audio (Part I)

Subjective evaluation of spatial distorsions induced by a sound source separation process
S. Fargeot, O. Derrien, G. Parseihian, M. Aramaki and R. Kronland-Martinet 67

EEG measurement of binaural sound immersion
R. Nicol, O. Dufor, L. Gros, P. Rueff and N. Farrugia 73

Binaural sound rendering improves immersion in a daily usage of a smartphone video game
J. Moreira, L. Gros, R. Nicol, I. Viaud-Delmon, C. Le Prado and S. Natkin 79

Discrimination experiment of sound distance perception for a real source in near-field
Z. Guo, Y. Lu, L. Wang and G. Yu 85

Auditory vertical localization in the median plane with conflicting spectral and dynamic cues
B. Xie, J. Jiang, C. Zhang and L. Liu 91

Towards a perceptually optimal bias factor for directional bias equalisation of binaural ambisonic rendering
T. McKenzie, D. Murphy and G. Kearney 97

Session P029 - Poster session

The SpHEAR project update: Refining the OctaSpHEAR, a 2nd order ambisonics microphone
F. Lopez-Lezcano 103

Spaces @ CCRMA: Design and evolution of our 3D studio and concert diffusion systems
F. Lopez-Lezcano 109

A very simple way to simulate the timbre of flutter echoes in spatial audio
T. Halmrast 115

Perceptual comparison of ambisonics-based reverberation methods in binaural listening
I. Engel, C. Henry, S.V. Amengual Garí, P.W. Robinson, D. Poirier-Quinot and L. Picinali 121

A linear phase IIR filterbank for the radial filters of ambisonic recordings
C. Langrenne, E. Bavu and A. Garcia 127

Investigation of sweet spot radius of sound reconstruction system based on inverse filtering
H. Okumura and M. Otani 133

Detecting the direction of emergency vehicle sirens with microphones
N.R. Shabtai and E. Tzirkel 137

Acoustic simulation of Bach's performing forces in the Thomaskirche
B. Boren, D. Abraham, R. Naressi, E. Grzyb, B. Lane and D. Merceruio 143

Session R014 - Perceptual evaluation of spatial audio (Part II)

An evaluation of pre-processing techniques for virtual loudspeaker binaural ambisonic rendering
T. McKenzie, D. Murphy and G. Kearney 149

Computational models for listener-specific predictions of spatial audio quality
P. Majdak and R. Baumgartner 155

Flexible binaural resynthesis of room impulse responses for augmented reality research
S.V. Amengual Garí, W.O. Brimijoin, H.G. Hassager and P.W. Robinson 161

Session R015 - Headphone and loudspeaker based reproduction

Multi-zone sound field reproduction via sparse spherical harmonic expansion
A. Dagar and R.M. Hegde 167

Parametric first-order ambisonic decoding for headphones utilising the cross-pattern coherence algorithm
L. McCormack and S. Delikaris-Manias 173

Inter-frequency band correlations in auditory filtered median plane HRTFs
Y. Iwaya, B.F.G. Katz, T. Magariyachi and Y. Suzuki 179

DEEP-SOUND FIELD ANALYSIS FOR UPSCALING AMBISONICS SIGNALS

Gyanajyoti Routray, Sourya Basu, Pranay Baldev, and Rajesh M Hegde
Department of Electrical Engineering
Indian Institute of Technology, Kanpur, India
{groutray,souryab, bpranay, rhegde}@iitk.ac.in

ABSTRACT

Higher Order Ambisonics (HOA) is a popular method for rendering spatial audio. However, the desired sound field can be reproduced over a small reproduction area at lower ambisonic orders. This problem can be handled by upscaling B-format signals using several methods both in the time and frequency domain. In this paper, a novel Sequential Multi Stage DNN (SMS-DNN) is developed for upscaling Ambisonic signals. The SMS-DNN allows for training of a very large number of layers since training is performed in blocks consisting of a fixed number of layers. Additionally, the vanishing gradient problem in DNN with a large number of layers is also effectively handled by the proposed SMS-DNN due to its sequential nature. This method does not require prior estimation of the source locations and works in multiple source scenarios. Reconstructed sound field analysis, subjective and objective evaluations conducted on the upscaled Ambisonic sound scenes indicate reasonable improvements when compared to the benchmark HOA reproduction.

1. INTRODUCTION

Spatial sound reproduction using Higher Order Ambisonics (HOA) is one of the most promising techniques for spatial audio reproduction [1, 2]. The knowledge of spherical harmonic decomposition is used to render spatial sound herein. But such a rendering has the limitation of low spatial resolution. Spatial resolution can be improved by increasing the number of loudspeakers during reproduction [3–5]. The preferred number of loudspeakers (L) is computed using the inequality $L \geq (N+1)^2$, N being the order of HOA [5]. Additionally, a large number of loudspeakers results in an under-determined system of equations. Consequently increase in number of loudspeakers is not a good choice [6]. However, it can be improved by the combined effort of upscaling the Ambisonic order and increasing the number of loudspeakers during sound reproduction. The upscaling can be done using compressed sensing technique [7] in time domain. Alternatively, in the frequency domain it can performed as in [8]. These are the sparsity based methods where the source and its direction are computed using an overcomplete spherical harmonics dictionary. The accuracy of this method depends on selection of the dictionary and estimation accuracy of the source location. These techniques have a limitation on the number of sources that can be rendered accurately. Real time upscaling can also be performed from multi channel recordings using spherical microphone array (SMA) such as Eigenmike® [9] and Visisonics® [10]. The number of microphones available and the design of spherical microphone array limits the order of HOA to 4 and 7 respectively.

In this work, a Sequential Multi Stage Deep Neural Network (SMS-DNN) for upscaling the order of Ambisonic signals is proposed and developed. The source sound is recorded using a tetrahedron microphone which gives a B-format (order-1 ambisonics) encoded signal [11]. The recordings are represented as four channels: W (omnidirectional) and (X, Y, Z) channels (bidirectional sounds in the direction of x, y, and z axes) respectively. These signals are upscaled into order-N HOA encoded plane wave sounds using the SMS-DNN in this work. Subsequently a complete framework is developed for upscaling Ambisonic signals using the SMS-DNN.

The rest of the paper is organized as follows. In Section 2 the problem for upscaling the lower order Ambisonics is described. The proposed SMS-DNN for upscaling Ambisonic signals is described in Section 3. The performance of the proposed method is evaluated in Section 4. Section 5 concludes the paper.

2. PROBLEM FORMULATION

A source vector consisting of P plane waves is given by

$$\mathbf{s} = [s_1, s_2, \ldots, s_P]$$

Where $(\theta_{s_1}, \phi_{s_1}), (\theta_{s_2}, \phi_{s_2}), \ldots, (\theta_{s_P}, \phi_{s_P})$, represent the location of the plane wave sources. Individually, θ represents the elevation, and ϕ the azimuth measured anticlockwise from $x-$axis. In general the HOA encoded signal is computed as [3]

$$\mathbf{B} \triangleq \mathbf{Y}\mathbf{s} \tag{1}$$

where $\mathbf{Y} = [Y_{nm}(\theta_{s_1}, \phi_{s_1}), \ldots, Y_{nm}(\theta_{s_P}, \phi_{s_P})]$ defines the spherical harmonics matrix, $n = 0, \ldots, N$ and $m =$

© Gyanajyoti Routray, Sourya Basu, Pranay Baldev, Rajesh M Hegde. Licensed under a Creative Commons Attribution 4.0 International License (CC BY 4.0). **Attribution:** Gyanajyoti Routray, Sourya Basu, Pranay Baldev, Rajesh M Hegde. "Deep-Sound Field Analysis for Upscaling Ambisonics Signals", 1st EAA Spatial Audio Signal Processing Symposium, Paris, France, 2019.

doi:10.25836/sasp.2019.14

$0, \ldots, n$ being the order and degree of spherical harmonic coefficients respectively. The spherical harmonic coefficients are defined as

$$Y_{nm}(\theta, \phi) = \frac{1}{2}\sqrt{\frac{(2n+1)(n-|m|)!}{\pi(n+|m|)!}} P_n^{|m|}(\cos\theta) e^{im\phi} \tag{2}$$

$P_n^{|m|}(\cdot)$ is defined as the normalized associated Legendre function of degree n and order m. From (2), it can be observed that every increase in order adds a pair of lobes. A simple way of decoding the encoded HOA is to place speakers in alignment with the direction of spherical harmonic functions and assign gains proportional to the directivity pattern of the source [6]. The sound field produced in such a decoding method is influenced by the interference width of the directivity pattern. For higher order Ambisonics the directive patterns are narrower, which results in improved spatial resolution. In the case of loudspeakers which are arranged uniformly in an icosahedron pattern on a sphere, the simplest way of obtaining the loudspeaker feeds (decoding) using HOA method is given by

$$\mathbf{g} = \mathbf{D}_{nm}\mathbf{B} \tag{3}$$

where $\mathbf{D}_{nm} = [Y_{nm}(\theta_{l_1}, \phi_{1_1}), \ldots, Y_{nm}(\theta_{l_L}, \phi_{l_L})]^\dagger$, L being the number of loud speakers satisfying $L \geq (N+1)^2$, and $\dagger$ representing the pseudo inverse of a matrix. Increasing the number of loudspeakers results in an undetermined system of equations. Hence for any improvement in the area of reproduction it is required to increase the ambisonic order. Due to the limitation on the number of microphones in spherical array processing, the order cannot be arbitrarily increased. In case of lower order ambisonic (B-format) audio recordings, the order can be upscaled only if the source direction is known. Additionally this method is limited to single source scenarios only. In this context the problem of upscaling can be modeled as a transfer $\mathfrak{F}(\cdot)$, that transfers the lower order spherical harmonics $Y_{nm}(\theta, \phi)|_{N=1}$ to the higher order spherical harmonics $Y_{nm}(\theta, \phi)|_{N>1}$. Due to the linear dependency between the HOA coefficients and the source signal the upscaling process can be defined as

$$Y_{nm}(\theta, \phi)|_{N=1}\mathbf{s} \xrightarrow{\mathfrak{F}} Y_{nm}(\theta, \phi)|_{N>1}\mathbf{s} \tag{4}$$

For multiple sources this can be formulated as

$$\mathbf{Y}_{nm}\mathbf{s} \xrightarrow{\mathfrak{F}} \mathbf{Y}_{nm}^{upscaled}\mathbf{s} \tag{5}$$

Where $\mathbf{Y}_{nm}$ is the spherical harmonic matrix corresponding to the location of sources. In-order to develop a flexible, scalable, and high resolution method for upscaling Ambisonic signals, a Sequential Multi Stage DNN (SMS-DNN) is proposed and developed in this work. The additional novelty of this method also lies in the fact that no prior estimation of source locations is required even in multiple source scenarios.

3. SMS-DNN FOR HOA ENCODING

The SMS-DNN consists of sequentially stacked DNNs, where each of the stacked DNNs upscales the order of the signal by 1. One of the most important properties of spherical harmonics is that the components of the signals are independent of each other. Additionally for a particular (θ, ϕ), increase in the order of signal only adds coefficients to the higher order Ambisonics, keeping the lower order Ambisonic coefficients unchanged. This property motivated the training of the DNNs independently for each order upscaling. Fig.1 shows the structure of the DNN for upscaling a signal from order 1 to order N.

3.1 Training the DNN

In this section, development of the dataset for training the DNN and subsequent upscaling is discussed. An algorithm is also detailed for the proposed method of upscaling Ambisonic signals.

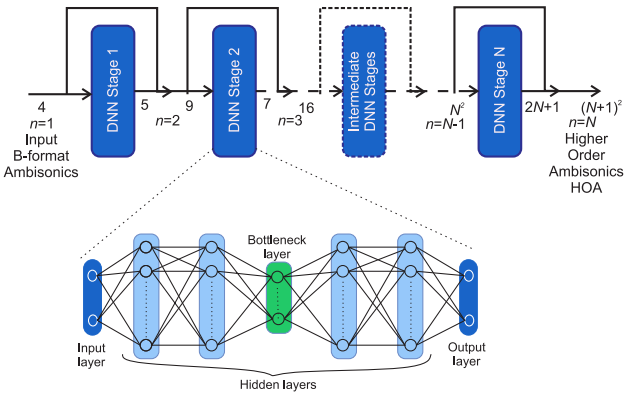


Figure 1. Sequential Multi Stage Deep Neural Network structure for upscaling Ambisonic signals.

3.1.1 Dataset for Training

The input training data of the deep neural network consists of an order N_l encoding of a mixture of sound signals located at K random locations (with random θ and ϕ). The output training data consists of a higher order encoding of the same mixture of sound sources with similar locations as the input data. For upscaling a signal from order N_l to N_u, we generate $N_u - N_l$ sets of training data, where the j^{th} dataset corresponds to the upscaling of the signal from order $N_l + j - 1$ to $N_l + j$. Additional details on the sequential training of the DNN is as follows.

3.1.2 Algorithm development for upscaling Ambisonic signals using SMS-DNN

Training a deep neural network for upscaling is a non-trivial problem, since the input to the DNN is a mixture of signals generated at different locations, hence there can exist multiple solutions for θ_s, ϕ_s and the amplitude of the sound sources given only the mixture. However, such solutions can be complex for a simple feed-forward network to compute. Also, we note that, for $N_u > N_l$, the order N_u

doi:10.25836/sasp.2019.14

signal of length $(N_u + 1)^2$ has the initial $(N_l + 1)^2$ values exactly equal to an order N_l signal. Further, note that the essential information that a DNN should extract from the input is a very small number of unknowns, such as location and amplitude parameters of the sound source for reproducing the entire higher order signal. The proposed SMS-DNN model for Ambisonic upscaling is capable of using a large number of layers without facing the typical problems such as vanishing gradient faced while increasing the number of layers in a single DNN [12]. The proposed method also uses a sequential approach to train the neural network. If the required upscaling task is from order N_l to N_u, then $N_u - N_l$ separate networks are trained independently. The j^{th} DNN is used to upscale from order $N_l + j - 1$ to $N_l + j$. Further, while upscaling a signal from order $N_l + j$ to $N_l + j + 1$, only the last $2(N_l + j) + 3$ entries of the upscaled values are required to be trained. Thus only $2(N_l + j) + 3$ output nodes are required in the j^{th} DNN, which makes the DNN fast and efficient even for large values of N_l and N_u. Finally, since the relevant information that needs to be extracted from the input layers is very small, a bottleneck hidden layer is introduced at the middle of the neural network with a smaller number of nodes [13, 14]. It was observed that using the bottleneck layer helps in faster convergence of the DNN compared to standard feed-forward DNN.

Algorithm 1: SMS-DNN for upscaling ambisonics signal to order N

Training:
Generate B-format dataset for K randomly located sound sources.
for $j = 1 : N - 1$ **do**
  Train the j^{th} DNN with order j sound signals as input and last $2j + 3$ elements of order $j + 1$ signal as desired output.
  Concatenate the $2j + 3$ output to the order j input signal to form the order $j + 1$ signal.
  The order $j + 1$ signal obtained from the concatenation is the input for the next DNN.
Return: Trained SMS-DNN.
Upscaling:
Input: Order 1 encoded signal of K sound sources.
for $j = 1 : N - 1$ **do**
  Feed the i^{th} DNN with the order j input and get $2j + 3$ output elements.
  Concatenate the $2j + 3$ output to the order j input signal to form the order $j + 1$ signal.
  The order $j + 1$ signal obtained from the concatenation is the input for the next DNN.
Return: Order N signal.

4. PERFORMANCE EVALUATION

Performance of the proposed method is evaluated using sound field reconstruction analysis, subjective and objective evaluations.

4.1 Experimental Conditions

The proposed method produces high resolution Ambisonics encoded signals from B-format encoded signals. Hence three sound scenes are created having five sounds each to evaluate the method. The scenes are created such that at any interval all the five sounds are overlapping. Each of the sound scenes are of length 10 to 15 seconds. B-format ambisonics signal of all the three sound scenes are upscaled to order 2 -order 7 using the SMS-DNN [1].

At the same time upscaled signal is also obtained using (1), which is referred to as benchmark encoded HOA signal. The mean square error between these two encoded signals is obtained and plotted in Figure 2. From the Figure 2, it is clear that the error propagates with the upscaling of ambisonic signals but it is bounded to $-5dB$.

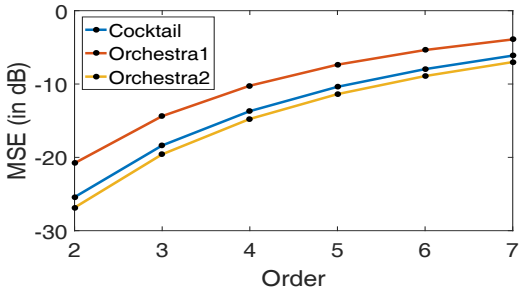


Figure 2. MSE between the reference and upscaled ambisonic encoded signals of various sound scenes for order 2 to 7.

4.1.1 Architecture of SMS-DNN

The proposed DNN for upscaling Ambisonic signals consists of a sequence of N-1 fully connected feed-forward neural networks, where each network is trained separately [15]. Each of the DNNs have 5 hidden layers, which consists of 300 nodes except the 3rd layer where only 20 nodes used to introduce bottleneck DNN structure.

4.1.2 Training Dataset

SMS-DNN was trained using 4×10^5 number of sample data points, where each of the training data is represented as a mixture of five randomly located sound sources. The elevation and azimuth were chosen randomly in the interval $\theta \in (0, \pi/2)$ and $\phi \in (0, \pi)$ respectively.

4.2 Analysis of Reconstructed Sound Field

A monochromatic source of frequency 2kHz, with a reproduction area of $0.64m^2$ centered around the receiver is considered. Spatially decoded signals are obtained using the conventional HOA decoder for the arrangement of loudspeaker as described in section 4.3. The sound density plots are shown in Figure 3 which illustrates improvement in spatial resolution as the ambisonics order increases. Average Error Distribution (AED) were calculated for the reproduced sound field using the proposed SMS-DNN by

[1] http://home.iitk.ac.in/~groutray/upscale_demo.html

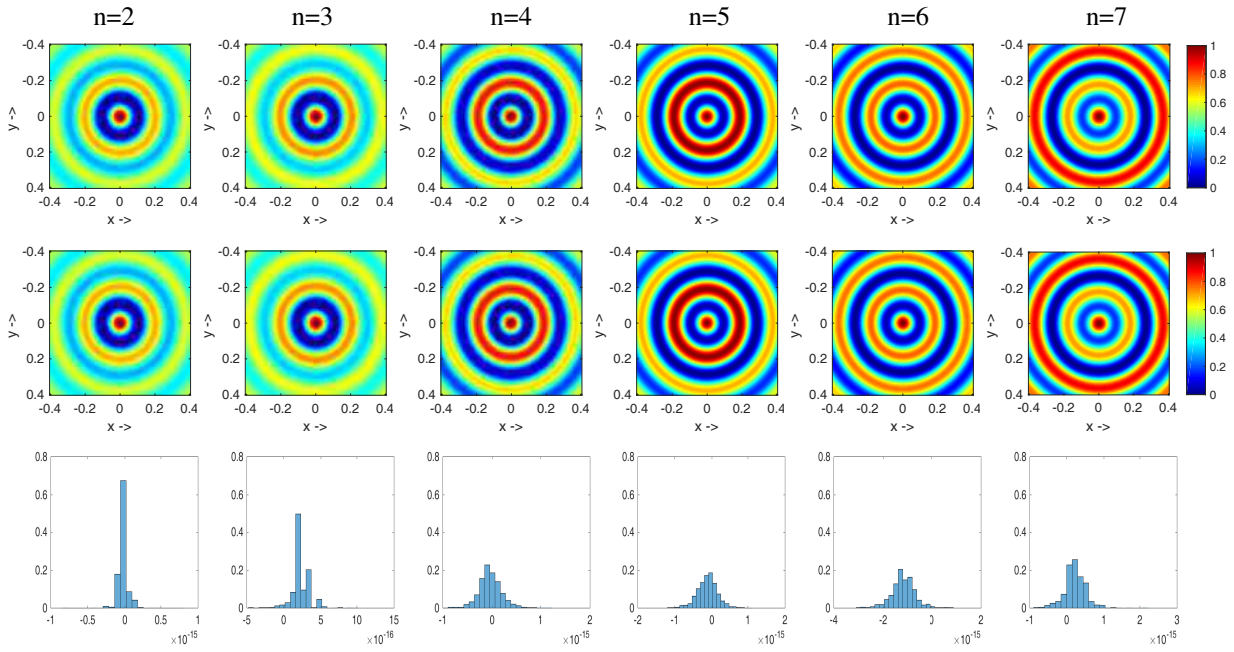


Figure 3. Sound pressure plots for order $n = 2$ to 7 (left - right) for a frequency 2kHz. Row-1 and row-2 represent the reference and SMS-DNN upscaled sound fields respectively. Row-3 illustrates Average Error Distribution for the reconstructed sound fields for order $n = 2$ to 7.

varying the order of n, from 2 to 7, as shown in 3rd row of Figure 3. From the Figure 3, it can be observed that the error of reconstruction reduces as the order of ambisonic is increased. Figure 3 shows significant improvement in the area of error free reproduction as the order increases.

4.3 Subjective and Objective Evaluations

For the analysis of spatial sound reproduction quality, a conventional HOA decoding procedure is adopted. For this 12, 24, 32, 42, 52, and 64 number of loudspeakers are used for orders from 2 to 7 respectively. The loudspeaker positions are found using the spherical t-design structure [16], such that the spherical harmonic matrix is well conditioned. Both subjective and objective evaluations were conducted for all the three sound scenes created earlier.

4.3.1 Subjective Evaluation

MUlti Stimulus test with Hidden Reference and Anchor (MUSHRA) [17] test is conducted for perceptual evaluation. First order Ambisonics is used as the anchor. The reproduced signal using the VisiSonics [18] spherical microphone array is used as the reference signal for the test. Fifteen participants were asked to rate the three scenes in a progressive scale of 0 for bad to 5 for excellent. The results of the MUSHRA test are shown in Figure 4. From Figure 4 it is observed that as the order of ambisonics increases the mean opinion scores for these three sound scenes also increases. The scores for the order 5 to 7 are clearly indicate an improvement in perceptual quality in comparison to the lower orders. The improvement is due to the fact that the area of perceptual reception of spatial audio increases as order increases. Hence it cam be anonymously conveyed that as order increases, the quality perception of the spatial audio also increases.

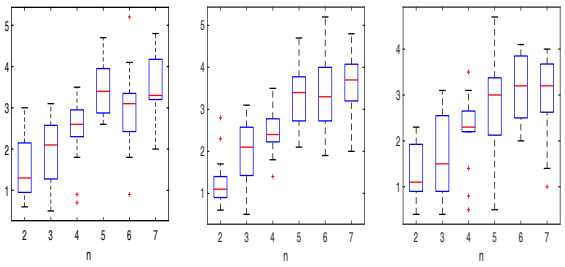


Figure 4. Mean perception score for various scenes (a) Cocktail (b) Orchestra1, and (c) Orchestra2.

4.3.2 Objective Evaluation

For objective evaluation, the methods as in [19, 20] are adopted. The perceptual quality of the sound was measured in terms of Perceptual Evaluation of Audio Quality (PEAQ) and Perceptual Similarity measure (PSM). The results for these tests are listed in Table 1. PEAQ is an objective measure in a scale 0 (for excellent) to -5 (for annoying). The measure Distortion index (DI) that compares the observed audio with the reference audio to find the distortion between the two audios. The absolute value closer to unity represents reduced distortion in the reproduced audio. PSM measures the similarity between the observed and reference audio. PSMt represents the fifth percentile of the sequence of instantaneous audio quality. In both the situation unity represents the perfect matching between the observed audio and reference audio. From the Table-1, it is observed that at higher orders, especially n=7

doi:10.25836/sasp.2019.14

for PEAQ and PSMt show performance score decreasing towards annoying, while subjective results show increasing performance as order is increasing. The objective performance tends to annoying as order increases due to the propagation of error. But the area of error free reproduction increases as the order increases. Hence the perceptual quality improves as the spatial resolution is improved due to upscaling.

Table 1. Objective evaluation scores for various sound scenes

N	Sound Scene	PEAQ	DI	PSM	PSMt
3	Cocktail	-1.8840	-	0.9630	0.8428
	Orchestra1	-2.5177	-0.1146	0.9577	0.7566
	Orchestra2	-2.1330	-	0.9781	0.8145
4	Cocktail	-2.5968	-1.0520	0.9312	0.7316
	Orchestra1	-3.0070	-1.3530	0.9232	0.6274
	Orchestra2	-2.7488	-1.0696	0.9559	0.6937
5	Cocktail	-2.9960	-1.9506	0.9002	0.6390
	Orchestra1	-3.2704	-2.0800	0.8923	0.5209
	Orchestra2	-3.0970	-1.9139	0.9317	0.5913
6	Cocktail	-3.2382	-2.5109	0.8674	0.5674
	Orchestra1	-3.4343	-2.5387	0.8643	0.4448
	Orchestra2	-3.3090	-2.4445	0.0982	0.5087
7	Cocktail	-3.3956	-2.8780	0.8428	0.5103
	Orchestra1	-3.5407	-2.8367	0.8440	0.3939
	Orchestra2	-3.4480	-2.8030	0.8889	0.4458

5. CONCLUSION

In this work a sequential multi stage DNN is proposed and developed for upscaling ambisonics signals. HOA encoded signal obtained from the trained SMS-DNN is compared with the reference HOA signal for evaluation. Analysis of sound field shows that the proposed upscaling technique improves the spatial resolution and reduces the error variance as observed from AED at varying higher ambisonic orders. This work can be extended to provide various perceptual quality improvements by infusing novelty into the training methodology followed. The extension of this approach to model based and parametric methods of spatial audio reproduction is currently been investigated.

6. ACKNOWLEDGEMENT

This work was funded by the SERB-DST under project no. SERB/EE/2017242.

7. REFERENCES

[1] J. Daniel, *Représentation de champs acoustiques, application à la transmission et à la reproduction de scènes sonores complexes dans un contexte multimédia*. PhD thesis, University of Paris VI, France, 2000.

[2] J. Daniel and S. Moreau, "Further study of sound field coding with higher order ambisonics," in *Audio Engineering Society Convention 116*, Audio Engineering Society, 2004.

[3] A. Wabnitz, N. Epain, A. van Schaik, and C. Jin, "Time domain reconstruction of spatial sound fields using compressed sensing," in *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, pp. 465–468, IEEE, 2011.

[4] D. Excell, "Reproduction of a 3d sound field using an array of loudspeakers," Master's thesis, 2003.

[5] D. B. Ward and T. D. Abhayapala, "Reproduction of a plane-wave sound field using an array of loudspeakers," *IEEE Transactions on speech and audio processing*, vol. 9, no. 6, pp. 697–707, 2001.

[6] S. Moreau, J. Daniel, and S. Bertet, "3d sound field recording with higher order ambisonics–objective measurements and validation of a 4th order spherical microphone," in *120th Convention of the AES*, pp. 20–23, 2006.

[7] A. Wabnitz, N. Epain, A. McEwan, and C. Jin, "Upscaling ambisonic sound scenes using compressed sensing techniques," in *Applications of Signal Processing to Audio and Acoustics (WASPAA), 2011 IEEE Workshop on*, pp. 1–4, IEEE, 2011.

[8] A. Wabnitz, N. Epain, and C. T. Jin, "A frequency-domain algorithm to upscale ambisonic sound scenes," in *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 385–388, March 2012.

[9] "The Eigenmike Microphone Array [online]." Available: `http://www.mhacoustics.com`.

[10] "The Visisonics Microphone Array [online]." Available: `https://visisonics.com`.

[11] P. G. Craven and M. A. Gerzon, "Coincident microphone simulation covering three dimensional space and yielding various directional outputs," Aug. 16 1977. US Patent 4,042,779.

[12] S. Hochreiter, "The vanishing gradient problem during learning recurrent neural nets and problem solutions," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 6, no. 02, pp. 107–116, 1998.

[13] B. Zhang, L. Xie, Y. Yuan, H. Ming, D. Huang, and M. Song, "Deep neural network derived bottleneck features for accurate audio classification," in *Multimedia & Expo Workshops (ICMEW), 2016 IEEE International Conference on*, pp. 1–6, IEEE, 2016.

[14] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, *et al.*, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal processing magazine*, vol. 29, no. 6, pp. 82–97, 2012.

[15] Y. Bengio *et al.*, "Learning deep architectures for ai," *Foundations and trends® in Machine Learning*, vol. 2, no. 1, pp. 1–127, 2009.

[16] N. Sloane, R. Hardin, and P. Cara, "Spherical designs in four dimensions," in *Information Theory Workshop, 2003. Proceedings. 2003 IEEE*, pp. 253–258, IEEE, 2003.

[17] I. Recommendation, "1534-1: Method for the subjective assessment of intermediate quality level of coding systems," *International Telecommunication Union*, 2003.

[18] C. B. Barley and J. B. Roach, "Surveillance camera with rapid shutter activation," Dec. 27 2012. US Patent App. 13/169,818.

[19] T. Thiede, W. C. Treurniet, R. Bitto, C. Schmidmer, T. Sporer, J. G. Beerends, and C. Colomes, "Peaq-the itu standard for objective measurement of perceived audio quality," *Journal of the Audio Engineering Society*, vol. 48, no. 1/2, pp. 3–29, 2000.

[20] R. Huber and B. Kollmeier, "Pemo-qa new method for objective audio quality assessment using a model of auditory perception," *IEEE Transactions on audio, speech, and language processing*, vol. 14, no. 6, pp. 1902–1911, 2006.

 doi:10.25836/sasp.2019.14

FIRST-ORDER AMBISONIC CODING WITH QUATERNION-BASED INTERPOLATION OF PCA ROTATION MATRICES

Pierre Mahé[1,2] **Stéphane Ragot**[2] **Sylvain Marchand**[1]

[1] L3i, Université de La Rochelle, France
[2] Orange Labs, Lannion, France
pierre.mahe@orange.com, stephane.ragot@orange.com, sylvain.marchand@univ-lr.fr

ABSTRACT

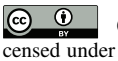
We present a new first-order ambisonic (FOA) coding method extending existing speech/audio codecs such as EVS or Opus. The proposed method is based on Principal Component Analysis (PCA) and multi-mono coding with adaptive bit allocation. The PCA operating in time-domain is interpreted as adaptive beamforming. To guarantee signal continuity between frames, beamforming matrices are interpolated in quaternion domain. The performance of the proposed method is compared with naive multi-mono coding with fixed bit allocation. Results show significant quality improvements at bit rates from 52.8 kbit/s (4×13.2) to 97.6 kbit/s (4×24.4) using the EVS codec.

1. INTRODUCTION

The current codecs used in telephony are mostly limited to mono. With the emergence of devices supporting spatial audio capture and playback, including multi-microphone smartphones, there is a need to extend traditional codecs to enable immersive communication conveying a spatial audio scene. There are different spatial audio coding approaches. For multichannel (or channel-based) audio, one can for instance use channel pairing with a stereo codec, parametric coding with downmixing/upmixing or residual coding. The most recent spatial audio codecs [1–4] handle various input types (channel, object, scene-based audio) and playback setups. Due to its flexibility, ambisonics is potentially an interesting internal coding representation to handle the multiplicity of input/output formats.

To code ambisonics, a naive approach is to extend an existing mono codec to spatial audio by coding each ambisonic component by separate codec instances; this approach is hereafter referred to as *multi-mono coding*. Informal listening tests showed that the naive multi-mono approach may create various spatial artifacts at low bit rates. These artifacts can be classified in three categories: diffuse blur, spatial centering, phantom source. They stem from the correlated nature of ambisonic components. A way to improve this is to use a fixed channel matrixing followed by multi-mono or multi-stereo coding as implemented in the ambisonic extension of Opus [5] or e-AAC+ [6]; this allows to better preserve the correlation structure after coding. Another approach is to analyze the audio scene to extract sources and spatial information. To code first-order ambisonic signal, the DirAC method [7] estimates the dominant source direction and diffuseness parameters in each time/frequency tile and re-creates the spatial image. This method has been extended to High-Order Ambisonics (HOA) in the so-called HO-DirAC [7] where the sound field is divided into angular sectors, for each angular sector, one source is extracted. More recently, Compass [8] was proposed: the number of sources is estimated and a beamforming matrix derived by Principal Component Analysis (PCA) is used to extract sources. In MPEG-H 3D Audio [1] a similar sound scene analysis is performed, e.g. by Singular Value Decomposition (SVD), to code ambisonics, and predominant and ambiance channels are extracted. The ambiance is transmitted as an FOA dowmix. When using PCA or SVD in time domain, transformed components may change dramatically between consecutive frames causing channel permutation and signal discontinuities [9]. The MPEG-H 3D Audio codec already employs overlap-add and channel re-alignent. In [9, 10], improvements were proposed, in particular performing SVD in frequency domain to ensure smooth transitions across frames.

In this work, we investigate how spatial audio can be coded by extending codecs currently used in telephony. We focus on FOA coding because this is a starting point before considering higher orders and FOA coding is required in some codec designs (e.g FOA truncation for ambiance [1] or HOA input in [11]). We reuse the Enhanced Voice Services (EVS) codec [12] which supports only mono input and output signals, however the proposed method can be applied to other codecs such as Opus. We wanted to avoid assumptions on the sound field (e.g. presence of predominant sources, number of sources). The aim was to use PCA to decorrelate FOA components prior to multimono coding; the PCA matrix can be seen as a matrix of beamformer weights. The continuity of the components across frames is guaranteed by an interpolation of 4D rotation ma-



© Pierre Mahé, Stéphane Ragot, Sylvain Marchand. Licensed under a Creative Commons Attribution 4.0 International License (CC BY 4.0). **Attribution:** Pierre Mahé, Stéphane Ragot, Sylvain Marchand. "First-Order Ambisonic Coding with Quaternion-Based Interpolation of PCA Rotation Matrices", 1st EAA Spatial Audio Signal Processing Symposium, Paris, France, 2019.

doi:10.25836/sasp.2019.19

trices in quaternion domain.

This paper is organized as follows. Section 2 gives an overview of ambisonics and beamforming. Section 3 provides some background on quaternions. Section 4 describes the proposed coding method. Section 5 presents subjective test results comparing the proposed method with naive multi-mono coding.

2. AMBISONICS

Ambisonics is based on a decomposition of the sound field in a basis of spherical harmonics. Initially limited to first order [13], the formalism was extended to high orders [14]. We refer to [15] for fundamentals of ambisonics. To perfectly reconstruct the sound field, an infinite order is required. In practice, the sound field representation is truncated to a finite order N. For an given order the number of ambisonic components is $n = (N+1)^2$.

In this work, we focus on first-order ambisonic (FOA) where $N = 1$ with $n = 4$ components (W, X, Y, Z). A plane wave with pressure $p(t)$ at azimuth θ and elevation ϕ (with the mathematical convention) is encoded to the following B-format representation:

$$\mathbf{y}(t) = \begin{bmatrix} w(t) \\ x(t) \\ y(t) \\ z(t) \end{bmatrix}^T = \begin{bmatrix} 1 \\ \cos\,\theta\,\cos\,\phi \\ \sin\,\theta\,\cos\,\phi \\ \sin\,\phi \end{bmatrix}^T p(t) \qquad (1)$$

To renderer ambisonics on loudspeakers, many decoding methods have been proposed – see for instance [16,17]. We only consider here the simplest method which recombines ambisonic components by a weight matrix $\mathbf{V} = [v_1, ..., v_n]$ to compute the signal feeds for loudspeakers located at known positions. This decoding is a conversion of the ambisonic representation to the loudspeaker domain. Similarly, it is possible to transform the ambisonic representation to another sound field representation using a transformation matrix $\mathbf{V}$ of size $n \times n$:

$$\mathbf{A} = \mathbf{V}\mathbf{Y} \qquad (2)$$

where $\mathbf{Y} = [\mathbf{y}_1, \cdots, \mathbf{y}_n]$ is the input matrix of n components, $\mathbf{A} = [\mathbf{a}_1, \cdots, \mathbf{a}_n]$ is the output matrix of the equivalent sound field representation. It is possible go from one sound field representation to another, provided that the matrix $\mathbf{V}$ is unitary:

$$\mathbf{V^T V} = \mathbf{I} \qquad (3)$$

where $\mathbf{I}$ is the identity matrix. If the matrix $\mathbf{V}$ does not satisfy this condition, some spatial deformation will occur.

It is possible to convert the representation $\mathbf{A}$ to the ambisonic representation $\mathbf{Y}$ by inverting the matrix $\mathbf{V}$. This principle of conversion from B-format to A-format is used in [6] or in the equivalent spatial domain (ESD) [18]. This conversion by a unitary matrix $\mathbf{V}$ allows interpreting the Principal Component Analysis (PCA) described hereafter, in terms of ambisonic transformation to an equivalent spatial representation.

To render ambisonics over headphones, a binaural rendering of ambisonic signals can be used. The simplest approach is to decode the signals over virtual loudspeakers, convolve the resulting feed signals by Head-Related Impulse Responses (HRIRs) and combine the results for each ear. A more optimized method using generic HRIRs in B-format domain is used in [14].

3. QUATERNIONS

Quaternions were introduced in 1843 by Hamilton [19] to generalize complex numbers. They have many applications in mathematics [20], computer graphics [21] and physics (aerospace, robotics, etc.) [22]. A quaternion q is defined as $q = a + b\mathbf{i} + c\mathbf{j} + d\mathbf{k}$, where a, b, c, d are real and $\mathbf{i}^2 = \mathbf{j}^2 = \mathbf{k}^2 = \mathbf{ijk} = -1$. Quaternions are often used as a parametrization of 3D rotations, especially for rotation interpolation. We recall that the set of 3D rotations can be mapped to the unit-norm quaternions under a one-to-two mapping [20–22], i.e. each 3D rotation matrix maps to two antipodal unit-norm quaternions: q and $-q$. Spherical linear interpolation (slerp) consists in the following principle [22]:

$$slerp(q_1, q_2, \gamma) = q_1(q_1^{-1}q_2)^\gamma \qquad (4)$$

where q_1 and q_2 are respectively the starting and ending quaternions and $0 \leq \gamma \leq 1$ is the interpolation factor. This is equivalent to [22]:

$$slerp(q_1, q_2, \gamma) = \frac{\sin((1-\gamma)\Omega)}{\sin(\Omega)} q_1 + \frac{\sin(\gamma\Omega)}{\sin(\Omega)} q_2 \qquad (5)$$

where $\Omega = \arccos(q_1 \cdot q_2)$ is the angle between q_1 and q_2 and $q_1 \cdot q_2$ is the dot product of q_1 and q_2. This boils down to interpolating along the grand circle (or geodesics) on a unit sphere in 4D with a constant angular speed as a function of γ. To ensure that the quaternion trajectory follows the shortest path on the sphere [21], the relative angle between successive unit-norm quaternions needs to be checked to choose between $\pm q_2$.

In this work, we used double quaternions to represent 4D rotation matrices. The product $\mathbf{R} = \mathbf{Q}^\star.\mathbf{P} = \mathbf{P}.\mathbf{Q}^\star$ of an anti-quaternion matrix $\mathbf{Q}^\star$ and a quaternion matrix $\mathbf{P}$, where

$$\mathbf{Q}^\star = \begin{pmatrix} a & b & c & d \\ -b & a & -d & c \\ -c & d & a & -b \\ -d & -c & b & a \end{pmatrix} \qquad (6)$$

and

$$\mathbf{P} = \begin{pmatrix} w & -x & -y & -z \\ x & w & -z & y \\ y & z & w & -x \\ z & -y & x & w \end{pmatrix} \qquad (7)$$

associated to $q = a + b\mathbf{i} + c\mathbf{j} + d\mathbf{k}$ and $p = w + x\mathbf{i} + y\mathbf{j} + z\mathbf{k}$, is a 4D rotation [20]. Conversely, given a 4D rotation $\mathbf{R}$ one may compute two quaternions q and p (up to sign) using factorization methods described in [20, 23]. The interpolation of 4D rotation matrices can be done by interpolating separately the associated pairs of quaternions, for instance using slerp interpolation. However, it is important to

doi:10.25836/sasp.2019.19

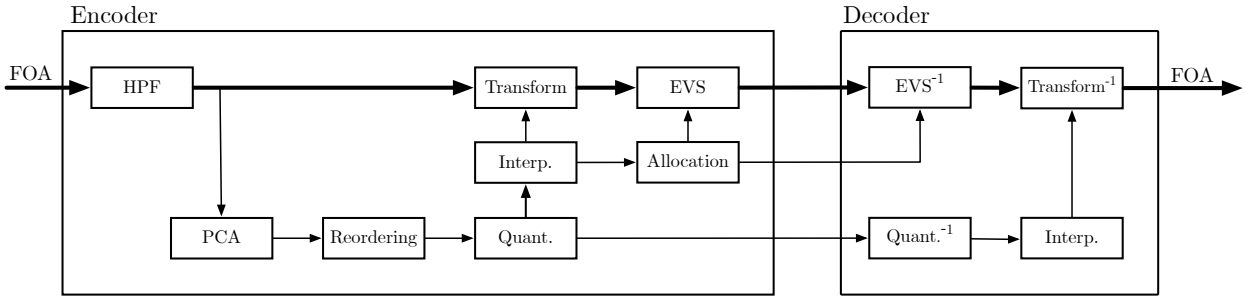


Figure 1. *Overview of proposed coding method.*

keep the sign consistent between double quaternions when constraining the shortest path.

4. PROPOSED CODING METHOD

The proposed coding method relies on a pre-processing of ambisonic components to decorrelate them prior to multi-mono coding. We illustrate this method using the EVS codec as a mono core codec. The input signal is a first-order ambisonic signal, with $n = 4$ ambisonic components labeled with an index $i = 1, ..., n$. The ambisonic channel ordering has no impact, therefore it is not specified here. The coding method operates on successive 20 ms frames, which is the EVS frame length. The overall codec architecture is shown in Figure 1. In the following we describe separately the pre-processing part based on PCA and the multi-mono coding part. We refer to [24] for more details on the codec description.

4.1 FOA pre-processing based on PCA

4.1.1 Beamforming matrix estimation

In each frame, the covariance matrix $\mathbf{C_{YY}}$ is estimated in time domain:

$$\mathbf{C_{YY}} = \mathbf{Y}^T\mathbf{Y} \tag{8}$$

where $\mathbf{Y} = [\mathbf{y}_1, ..., \mathbf{y}_n]$ is the matrix of $n = 4$ ambisonic components. The covariance matrix $\mathbf{C_{YY}}$ is factorized by eigenvalue decomposition as:

$$\mathbf{C_{YY}} = \mathbf{V}\Lambda\mathbf{V}^T \tag{9}$$

The matrix $\mathbf{V} = [\mathbf{v}_1, ..., \mathbf{v}_n]$ is a 4D rotation matrix if the transformation matrix is orthogonal and if $\det(\mathbf{V}) = +1$. We ensure that the eigenvector matrix defines a rotation matrix by inverting the sign of $\mathbf{v}_n$ if $\det(\mathbf{V}) = -1$. This matrix $\mathbf{V}$ is the transformation matrix to convert the components to another spatial domain equivalent to the original ambisonic B-format. In the following, the rotation matrix in the current frame of index t will be denoted $\mathbf{V}_t$.

To avoid bias from low frequencies in PCA, the input components are pre-filtered by a 20 Hz high-pass IIR filter from the EVS codec [25].

4.1.2 Re-alignment of beams

From frame to frame, the eigen decomposition can switch the order of eigenvectors or invert their sign. These changes may modify significantly the weighting coefficients of the beamforming matrix. Therefore the beam directions might change and these modifications might create discontinuities in signals, which can degrade audio quality after multi-mono coding. To improve signal continuity between successive frames, a signed permutation was applied to the eigenvector matrix in the current frame $\mathbf{V}_t$ to maximize similarity to the eigenvector matrix $\mathbf{V}_{t-1}$. The signed permutation is obtained in two steps:

First, a permutation is found to match the eigenvectors of frames t and $t-1$. This problem is treated as an assignment problem where the goal is to find the closest beam, in terms of direction. As in [9], the Hungarian algorithm was used with the following optimization criterion:

$$\mathbf{J}_t = tr(|\mathbf{V}_t.\mathbf{V}_{t-1}^T|) \tag{10}$$

Where $tr(|.|)$ is the trace of the matrix $|\mathbf{V}_t.\mathbf{V}_{t-1}^T|$ whose coefficients are the absolute values. It is noted that only the beams direction it is considered in this step.

Second, to avoid sign inversion of component between consecutive frames, the autocorrelation was computed:

$$\mathbf{\Gamma}_t = \tilde{\mathbf{V}}_t.\mathbf{V}_{t-1}^T \tag{11}$$

A negative diagonal value in $\mathbf{\Gamma}_t$ indicates a sign inversion between two frames. The sign of the respective columns of $\tilde{\mathbf{V}}_t$ was inverted to compensate for this change of direction.

4.1.3 Quantization of beamforming matrix

In [26], 2D and 3D rotation matrices were converted by angle parameters. A similar idea was used to quantize the beamforming matrix in each frame. A rotation matrix of size $n \times n$ can be parametrized by $n(n-1)/2$ generalized Euler angles [27]. The 4D rotation matrix $\mathbf{V}_t$ is converted to 6 generalized Euler angles, with 3 angles in $[-\pi, \pi)$ and 3 angles in $[-2\pi, 2\pi)$. These angles are coded by scalar quantization with a budget of respectively 8 and 9 bits for angles defined over a support of length π and 2π. The overall budget for 6 angles is 51 bits per frame.

4.1.4 Interpolation of beamforming matrices

To improve continuity and guarantee smooth transition between beams across consecutive frames, the 4D rotation matrix of the current frame and previous frames are interpolated by subframes. The rotation matrices are converted to pairs of quaternions (q, p) and the interpolation is done

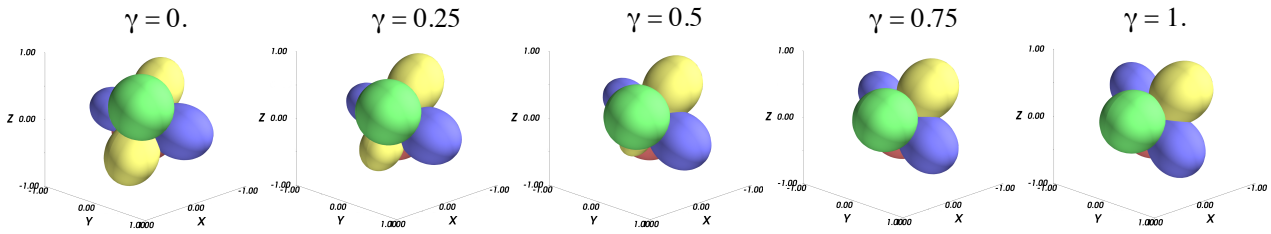


Figure 2. *Beamforming interpolation.*

in this double quaternion domain, in aim to interpolation of rotation with constant angular velocity. Each frame is divided into K subframes and for each subframe of index $1 \leq k \leq K$ in the current frame, the corresponding quaternion pairs (q_{t-1}, q_t) and (p_{t-1}, p_t) are interpolated used spherical linear interpolation (slerp). In the proposed coding method, the frame of $L = 640$ samples (20 ms at the 32 kHz sampling rate) is divided into K sub-frames. We used $K = 128$ which gives a subframe length of $L/K = 10$ samples (0.3125 ms) and the interpolation factor in Eq. 5 was set to $\gamma = k/K$. The interpolated pairs of quaternions were converted back to a 4D matrix.

The beamforming matrix interpolation is illustrated in Figure 2, for interpolation factors $\gamma = 0, 0.25, 0.5, 0.75$ and 1.

4.1.5 PCA matrixing

The pre-processed FOA signal is transformed into 4 principal components by applying the interpolated 4D rotation matrix (beamforming matrix) in each sub-frame.

4.2 Multi-mono coding with adaptive bit rate allocation

In naive multi-mono coding the bit rate is the same for each component. It was observed experimentally that signal energy after PCA matrixing may vary significantly between components and it was found that an adaptive bit allocation is necessary to optimize quality. The EVS codec quality [28] does not increase according to rate-distortion theoretic predictions when increasing bit rate. In this work the audio quality for each bit rate was modeled by energy-weighted average MOS (Mean Opinion Score) values. We used a greedy bit allocation algorithm which aims at maximizing the following score:

$$S(b_1, \cdots, b_n) = \sum_{i=1}^{n} Q(b_i).E_i^{\beta} \qquad (12)$$

where b_i and E_i are respectively the bit allocation and the energy of the i^{th} component in the current frame and $Q(b_i)$ is a MOS score for a bit rate corresponding to b_i bits. These values $Q(b_i)$ may be take from the EVS characterization report [28] the values used in this work are defined in [24]. This optimization is subject to the constraint $b_1 + \cdots + b_n \leq B$ where B is the budget allocated for multi-mono coding. Note that if another core codec than EVS is used, the values $Q(b_i)$ can be adjusted accordingly; for instance, a quality evaluation of Opus can be found in [29]. The bit allocation to individual audio channels was restricted to all EVS bit rates $\geq$ 9.6 kbit/s to ensure a super-wideband coded bandwidth. Details for bitstream structure and a bit rate allocation example can be found in [24].

In each 20 ms frame, the selected bit allocation is transmitted to the decoder and used for multi-mono EVS coding.

4.3 Decoding

The decoding part consists in multi-mono decoding based on the received bit allocation and PCA post-processing (which is the inverse of the pre-processing) in each frame.

5. EXPERIMENTAL RESULTS

5.1 Test setup

We conducted a subjective test according to the MUSHRA methodology [30] to compare the performance of naive multi-mono coding and the proposed coding method. For each item, subjects were asked to evaluate the quality of conditions with a grading scale ranging of 0 to 100 (Bad to Excellent). The test conditions included three specific items: the hidden reference (FOA) and two anchors. MUSHRA tests for mono signals typically use a low anchor (3.5kHz low-pass filtered original) and a medium anchor (7kHz low-pass filtered original). For MUSHRA tests with stereo, it is suggested to use a "reduced stereo image" as anchors [30]. There is no clear recommendation for spatial alterations for MUSHRA tests with ambisonics. In this work we used the following spatial reduction:

$$FOA = \begin{pmatrix} W \\ \alpha X \\ \alpha Y \\ \alpha Z \end{pmatrix}, \qquad \alpha \in [0, 1] \qquad (13)$$

with $\alpha = 0.65$ and $\alpha = 0.8$ for the low and medium anchors, respectively. All FOA items were binauralized with the Resonance Audio renderer [31]. All test conditions are summarized in Table 1. The test items consisted of 10 challenging ambisonic items: 4 voice items, 4 music items and 2 ambient scenes. The synthetic items were generated in Orange Labs, the recorded items were captured and mixed by Orange Labs or done jointly with partners, see [24] for more details. All subjects conducted the listening test with the same professional audio hardware in a dedicated listening room at Orange Labs. In total 11 listeners participated

doi:10.25836/sasp.2019.19

Table 1. *List of MUSHRA conditions.*

Short name	Description
HREF	FOA hidden reference
LOW_ANCHOR	3.5 kHz LP-filtered and spatially-reduced FOA ($\alpha = 0.65$)
MED_ANCHOR	7 kHz LP-filtered and spatially-reduced FOA ($\alpha = 0.8$)
MULTI52	FOA coded by multimono EVS at 4×13.2 kbit/s
MULTI65	FOA coded by multimono EVS at 4×16.4 kbit/s
MULTI97	FOA coded by multimono EVS at 4×24.4 kbit/s
PCA52	FOA coded by proposed method at 52.8 kbit/s
PCA65	FOA coded by proposed method at 65.6 kbit/s
PCA97	FOA coded by proposed method at 97.6 kbit/s

in the test; all of them are expert or experienced listeners without hearing impairments. Each item was coded at three bit rates for multi-mono coding: 52.8, 65.6, 97.6 kbit/s which corresponds to a fixed allocation of 13.2, 16.4 and 24.4 kbit/s per channel (respectively). For the proposed coding method, as explained in Section 4, the bit rate was dynamically distributed between channels; however the target (maximum) bitrate was set to the same bit rate as multi-mono coding for a fair comparison.

5.2 Subjective test results

The MUSHRA test results, including the mean and 95% confidence intervals, are presented in Figure 3. They show that significant quality improvement over multi-mono coding. For instance, the proposed coding method at 52.8 kbit/s is is equivalent to multi-mono coding at 65.6 kbit/s.

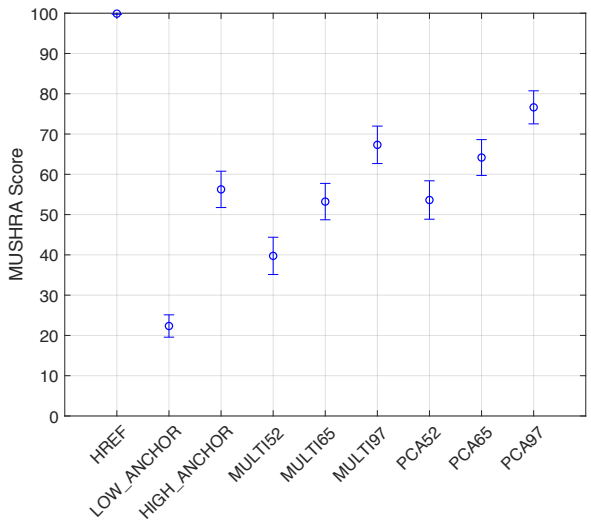


Figure 3. *MUSHRA test results.*

Spatial artifacts were noted at every bit rate for multi-mono coded items. They can be classified in three categories: diffuse blur, spatial centering, phantom source. With the proposed coding method, these artifacts are mostly removed because the correlation structure is less impacted by coding. This explanation was supported by the feedback from some subjects, after they conducted the subjective test.

6. CONCLUSION

This article presented a spatial extension of an existing codec (EVS), with a pre-processing decorrelating ambisonic components prior to multi-mono coding. The proposed method operate in time domain to avoid extra delay and allow maximum compatibility with existing codecs which are used as a black box. In each frame, a beamforming basis is found by PCA; the PCA matrices are interpolated in quaternion domain to guarantee smooth transitions between beamforming coefficients. Subjective test results showed significant improvements over naive multi-mono EVS coding for bit rates from 4×13.2 to 4×24.4 kbit/s, which may be explained by the combination of the use of PCA matrixing and adaptive bit allocation.

7. ACKNOWLEDGMENTS

The authors thank all participants in the subjective test. They also thank Jérôme Daniel for discussions on spatial reduction for MUSHRA anchors items.

8. REFERENCES

[1] J. Herre, J. Hilpert, A. Kuntz, and J. Plogsties, "MPEG-H audio—the new standard for universal spatial/3D audio coding," *Journal of the Audio Engineering Society*, vol. 62, no. 12, pp. 821–830, 2015.

[2] ETSI TS 103 190 V1.1.1, "Digital Audio Compression (AC-4) Standard," April 2014.

[3] ATSC Standard, Doc. A/52:2018, "Digital Audio Compression (AC-3, E-AC-3)," January 2018.

[4] ETSI TS 103 491 V1.1.1, "DTS-UHD Audio Format; Delivery of Channels, Objects and Ambisonic Sound Fields," April 2017.

[5] J. Skoglund, "Ambisonics in an Ogg Opus Container." IETF RFC 8486, Oct. 2018.

[6] 3GPP TS 26.918, "Virtual Reality (VR) media services over 3GPP, clause 6.1.6," 2018.

[7] V. Pulkki, A. Politis, M.-V. Laitinen, J. Vilkamo, and J. Ahonen, "First-order directional audio coding (DirAC)," in *Parametric Time-Frequency Domain Spatial Audio*, ch. 5, John Wiley & Sons, 2018.

[8] A. Politis, S. Tervo, and V. Pulkki, "Compass: Coding and multidirectional parameterization of ambisonic sound scenes," in *Proc. ICASSP*, pp. 6802–6806, 2018.

[9] S. Zamani, T. Nanjundaswamy, and K. Rose, "Frequency domain singular value decomposition for efficient spatial audio coding," in *Proc. WASPAA*, pp. 126–130, 2017.

[10] S. Zamani and K. Rose, "Spatial Audio Coding with Backward-Adaptive Singular Value Decomposition," in *145th AES Convention*, 2018.

[11] D. McGrath *et al.*, "Immersive Audio Coding for Virtual Reality Using a Metadata-assisted Extension of the 3GPP EVS Codec," in *Proc. ICASSP*, May 2019.

[12] S. Bruhn *et al.*, "Standardization of the new 3gpp evs codec," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5703–5707, IEEE, 2015.

[13] M. A. Gerzon, "Periphony: With-height sound reproduction," *Journal of the Audio Engineering Society*, vol. 21, no. 1, pp. 2–10, 1973.

[14] J. Daniel, *Représentation de champs acoustiques, application à la transmission et à la reproduction de scènes sonores complexes dans un contexte multimédia*. PhD thesis, Université Paris 6, 2000. `http://gyronymo.free.fr`.

[15] B. Rafaely, *Fundamentals of spherical array processing*. Springer, 2015.

[16] A. J. Heller, R. Lee, and E. M. Benjamin, "Is My Decoder Ambisonic?," in *125th AES Convention*, 2008.

[17] F. Zotter and M. Frank, "All-round ambisonic panning and decoding," *Journal of the Audio Engineering Society*, vol. 60, no. 10, pp. 807—820, 2012.

[18] 3GPP TS 26.260, "Objective test methodologies for the evaluation of immersive audio systems," 2018.

[19] W. R. Hamilton, *On a new Species of Imaginary Quantities connected with a theory of Quaternions*, vol. 2. 1844.

[20] P. De Casteljau, *Les quaternions*. Dunod, 1987.

[21] K. Shoemake, "Animating rotation with quaternion curves," *ACM SIGGRAPH Computer Graphics*, vol. 19, no. 3, pp. 245—-254, 1985.

[22] A. Hanson, *Visualizing Quaternions*. Morgan Kaufmann Publishers, 2006.

[23] A. Perez-Gracia and F. Thomas, "On Cayley's factorization of 4D rotations and applications," *Advances in Applied Clifford Algebras*, vol. 27, no. 1, pp. 523–538, 2017.

[24] P. Mahé, S. Ragot, and S. Marchand, "First-Order Ambisonic Coding with PCA Matrixing and Quaternion-Based Interpolation," in *Proc. DAFx*, 2019.

[25] 3GPP TS 26.445, "Codec for Enhanced Voice Services (EVS); Detailed algorithmic description," 2019.

[26] M. Briand, *Études d'algorithmes d'extraction des informations de spatialisation sonore : application aux formats multicanaux*. PhD thesis, INPG Grenoble, 2007.

[27] D. K. Hoffman, R. C. Raffenetti, and K. Ruedenberg, "Generalization of Euler Angles to N-Dimensional Orthogonal Matrices," *Journal of Mathematical Physics*, vol. 13, no. 4, pp. 528–533, 1972.

[28] 3GPP TS 26.952, "Codec for Enhanced Voice Services (EVS); Performance Characterization," 2019.

[29] A. Rämö and H. Toukomaa, "Voice quality characterization of IETF Opus codec," in *Proc. Twelfth Annual Conference of the International Speech Communication Association*, 2011.

[30] ITU-R Rec. BS.1534–3, "Method for the subjective assessment of intermediate quality level of coding systems," 2015.

[31] "Resonance audio : Rich, immersive, audio." `https://resonance-audio.github.io/resonance-audio`.

 doi:10.25836/sasp.2019.19

BLOCK-SPARSE APPROACH FOR THE IDENTIFICATION OF COMPLEX SOUND SOURCES IN A ROOM

Hugo Demontis[1] **Francois Ollivier**[1] **Jacques Marchal**[1]

[1] Sorbonne Université, CNRS, Institut Jean Le Rond dAlembert, UMR 7190, F-75005 Paris, France

hugo.demontis@sorbonne-universite.fr

ABSTRACT

This paper describes a practical methodology to estimate the directivity pattern of complex sound sources in reverberant environments. The acoustic analysis takes place under non-anechoic conditions. In this context, classical beamforming techniques fails to achieve this task because they generally rely on the free-field propagation model. In order to identify the directivity pattern with respect to frequency, an adapted block-sparse algorithm is proposed. This considers the first reflections to approximate the propagation path in the room, useful to solve accurately the related inverse problem. In addition, a large 3D array of microphones is implemented and validated. Largeness concerns here in both its dimensions and the number of microphones. The array consists of five planar sub-arrays which surround an entire room where sources are located. The microphones are flush mounted on the walls. From the pressure signal emitted by the source and recorded by the 1024 synchronous digital MEMS, a spherical harmonics representation of the source is then computed. A prototype of quadripolar source is tested to assert the efficiency of the proposed framework.

1. INTRODUCTION

From virtual reality to noise control, a wide range of applications in acoustics need the *directivity pattern* (DP) of a sound source to be known with a good approximation. To this end, surrounding spherical microphones arrays provide the most natural way to measure the DP experimentally [1]. In this type of antenna, sensors at fixed distance and distributed all around – i.e. at a solid angle of 4π – capture the direct sound field emitted by the complex sound source. For example, Behler proposes an arrangement of 32 microphones mounted on the vertices of a truncated icosahedron to evaluate the radiation nature of musical instruments. [2]. Spherical microphones array also improves the use of algorithms based on *spherical harmonics* (SH) formulation. The full 3-D pressure field representing the DP is then transformed to a 1-D vector containing few coefficients. The array design becomes here a relevant point, regarding the radius of the sphere, the number of microphones or their location along the spherical surface. This ensure to discard aliasing effects in the SH spectrum or to analyse the sound field along an extended frequency bandwidth. However, conventional recording systems generally suffer from low resolution in the SH domain, in part due to the small number of microphones used.

The DP estimation generally relies on a free field propagation model of the source. Most of the measurements are thus performed in an anechoic chamber. But under real conditions, like most of the time in confined spaces, specular reflections on the walls degrade the array signal. The performance of SH algorithms then decreases with the reverberation. In theory, considering the *Room Transfer Function* (RTF) to solve the inverse problem discard the room effects. But measuring a complete set of RTFs all over the volume is however unachievable due to the overwhelming number of samples needed to satisfy the Shannon sampling theorem [3].

The work in this paper presents the implementation of a complete methodology to overcome these two drawbacks. The numerical framework relies on the *Block Orthogonal Matching Pursuit* (BOMP) algorithm, which exploits the sparsity nature of the sources in the spatial domain [4]. The blocks refers in our case to the SH representation of the DP. This allows to joint the localization of the sources and their DP identification in a unique numerical task. The free-field version of the BOMP is then modified to be effective in reverberant environment, by including a closed form of the RTF based on the *Image-Source Method* (ISM) [5]. Only the early part is approximate here, involving the early specular reflections of the RTF. Beside we use a large 3D array with several hundred digital MEMS microphones. For practical reasons, the microphones are flush mounted on the walls and the ceiling of a rectangular room. Their location is known. Large aperture microphone arrays have already been studied in [6] and remains useful for sound source analysis along extended acoustic area.

This paper takes place in three parts. First, we establish the theory to understand the acoustic inverse problem regarding the estimation of the DP of sources. The BOMP algorithm is here introduced and its implementation is described. Secondly, we present the chosen acquisition system and the array deployment inside the room. Finally, we

© Hugo Demontis, Francois Ollivier, Jacques Marchal. Licensed under a Creative Commons Attribution 4.0 International License (CC BY 4.0). **Attribution:** Hugo Demontis, Francois Ollivier, Jacques Marchal. "Block-sparse approach for the identification of complex sound sources in a room", 1st EAA Spatial Audio Signal Processing Symposium, Paris, France, 2019.

doi:10.25836/sasp.2019.17

demonstrate the efficiency of the proposed method with an experiment. The SH nature of a quadripolar source, made with two unbaffled loudspeaker, is evaluated. Results are compared to the one obtained from previous work, involving a semicircular microphones array.

2. IDENTIFICATION FRAMEWORK

2.1 Source radiation model

Let a complex source at the origin of a 3D space described by the spherical coordinates $\mathbf{r} = (r, \Omega)$, with $\Omega = (\theta, \phi)$ i.e. the azimutal and zenithal couple. In free-field conditions, the continuous pressure distribution can be described at any point using the following SH expansion :

$$p(r, \Omega, k) = \sum_{l=0}^{L} \sum_{q=-l}^{l} \alpha_{lq}(k) h_l(kr) Y_l^q(\Omega) \qquad (1)$$

at the angular frequency $\omega = kc$, with wave number k and sound velocity c. The Hankel function of the second kind h_l refers to the radial acoustic spreading. The SH function Y_l^q of order l and degree q links with the angular dependency of the DP. More information about these mathematical objects can be found in [7]. The coefficients α_{lq} constitute the unknown in the DP estimation problem. They give the contribution of each SH to the pressure field and can be computed by the inverse transform of (1) :

$$\alpha_{lq}(k) = \frac{1}{h_l(kr)} \int_0^{2\pi} \int_0^{\pi} p(r, \Omega, k) Y_l^q(\Omega)^* d\Omega \qquad (2)$$

with $(\cdot)^*$ the conjugate operator. Note here that the pressure field is assumed to be band-limited up to the order L :

$$\alpha_{\tilde{L}q} = 0 \quad \forall \tilde{L} > L \qquad (3)$$

2.2 Discrete Formulation

The harmonic pressure field is sampled using M microphones. The signal p_m at the m^{th} microphone follows the free-field model in (1). The observations are stored in the vector $\mathbf{p}$ of size $M \times 1$. Sources are searched according to N grid points in a given volume of interest. For one at the n^{th} point, the matrix formulation of (2) writes :

$$\boldsymbol{\alpha}_n = \mathbf{H}_n^{\dagger} \mathbf{p} \qquad (4)$$

with $(\cdot)^{\dagger}$ the generalized inverse operator (or Moore-Penrose pseudoinverse). The set $\boldsymbol{\alpha}$ of size $(L+1)^2 \times 1$, containing the harmonic SH coefficients, is evaluated in a least-square sense. The steering matrix $\mathbf{H}_n$ of size $M \times (L+1)^2$ describes the propagation path from the n^{th} point-source to each microphone :

$$\mathbf{H}_n = [(\mathbf{h}_0^0)_n \ \ldots \ (\mathbf{h}_l^q)_n \ \ldots \ (\mathbf{h}_L^L)_n] \qquad (5)$$

The "atoms" $(\mathbf{h}_l^q)_n$

$$(\mathbf{h}_l^q)_n = [h_l(kr_{1n}) Y_l^q(\Omega_{1n}) \ \ldots \ h_l(kr_{Mn}) Y_l^q(\Omega_{Mn})]^{\mathrm{T}} \qquad (6)$$

of size $M \times 1$ depends on the spherical coordinates (r_{mn}, Ω_{mn}) i.e. the location of the m^{th} microphone regarding the n^{th} point.

2.3 Block-sparse approach

For a number of grid points greater than the number of microphones ($M \ll N$), this discrete inverse problem is under-determined and an infinite number of solutions exists. However, it can be regularized by exploiting the sparsity nature of the sources in the spatial domain ($S \ll N$). This assumption yields to approximate $\mathbf{p}$ by a linear combination of only S blocks of $\mathbf{H} = \{\mathbf{H}_n\}_{n=1:N}$. In this case, greedy algorithms like the BOMP remains a useful method to solve it. While the algorithm dedicates to recover a signal with a small number of measurements, our goal here is to jointly localize and identify multiple sources in the same numerical scheme. The n^{th} selected block refers to the s^{th} source and its harmonic SH coefficients are then estimated.

The framework proposed in this work derivates from the BOMP. It comprises four successive steps exploiting the discrete formulation in (4). Let $\mathbf{r}^0 = \mathbf{p}$ and $\mathbf{S} = \emptyset$:

1. Selecte the most correlated block $\mathbf{H}_n$ with the pressure vector $\mathbf{p}$:

$$s = \arg\max_n \left(\|\mathbf{H}_n^{\dagger} \mathbf{r}^{(i-1)}\|_{\infty} \right)$$

2. Update the set J with the index j :

$$\mathbf{S}^{(i)} = \mathbf{S}^{(i-1)} \cup s$$

3. Compute the orthogonal projector $\mathbf{\Pi}_{J^{(i)}}$ built from the selected blocks :

$$\mathbf{\Pi}_{\mathbf{S}^{(i)}} = \mathbf{H}_{\mathbf{S}^{(i)}} \mathbf{H}_{\mathbf{S}^{(i)}}^{\dagger}$$

4. Update the residual $\mathbf{r}^{(i)}$:

$$\mathbf{r}^{(i)} = (\mathbb{1} - \mathbf{\Pi}_{J^{(i)}}) \mathbf{r}^{(i-1)}$$

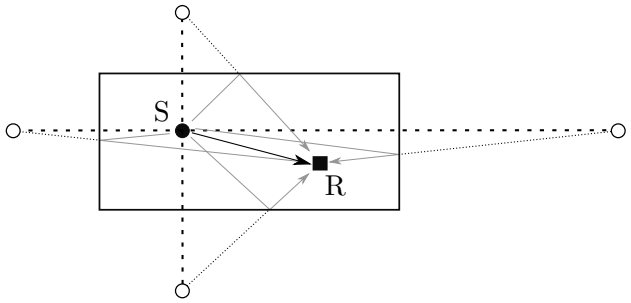
In case of S sources, these steps iterate until all the sources are discovered. Finally, the solution $\boldsymbol{\alpha}_{\mathbf{S}}$ is given as a matrix of size $(L+1)^2 \times S$:

$$\boldsymbol{\alpha}_{\mathbf{S}} = \mathbf{H}_{\mathbf{S}}^{\dagger} \mathbf{p} \qquad (7)$$

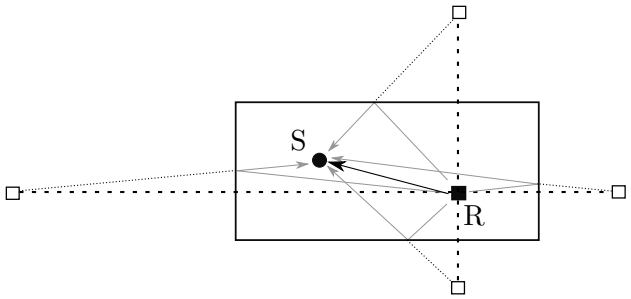
2.4 Under reverberant conditions

The steering matrix in (5) is only valid under anechoic conditions. Yet, algorithms based on free-field model tends to fail in rooms, when diffuse field interferes with the sources. This can be raised by including the RTF into the inverse problem solving. The RTF describes the propagation path between two points in an enclosed space. It takes into account the direct path and the multiple reflections caused by rigid boundaries and obstacles. For rooms perfectly rectangular, the ISM provides an efficient way to approximate it. The reflections are replaced by the free-field contributions from virtual sources (VS), located outside the walls. The real domain then extends to an infinity of virtual symmetric rooms. For each one mirrored, a virtual microphone array (VMA) can also be considered, recording the direct part

doi:10.25836/sasp.2019.17



(a) ISM



(b) IMM

Figure 1: Direct (ISM) and inverse (IMM) representation of the RTF.

of the associated VS. The total number J of VMAs relies directly with the order of reflections R. See [8] for a similar adaptation with classical delay-and-sum beamforming algorithm.

Applying the BOMP under reverberant conditions relies on this concept With the *Image-Microphone Method* (IMM), the process is performed not only with the real antenna, but also with all the considered VMAs. The implementation of the IMM stills straightforward. The computation of the parameters, like the reflection coefficients β_j or the spherical coordinates $(r^j_{mn}, \Omega^j_{mn})$ of the j^{th} VMA, is analog to those for the ISM.

2.5 Array description

The steering matrix in (5) is only valid under anechoic conditions. Yet, algorithms based on free-field model tends to fail in rooms, when diffuse field interferes with the sources. This can be raised by including the RTF into the inverse problem solving. The RTF describes the propagation path between two points in an enclosed space. It takes into account the direct path and the multiple reflections caused by rigid boundaries and obstacles. For rooms perfectly rectangular, the ISM provides an efficient way to approximate it. The reflections are replaced by the free-field contributions from virtual sources (VS), located outside the walls. The real domain then extends to an infinity of virtual symmetric rooms. For each one mirrored, a virtual microphone array (VMA) can also be considered, recording the direct part of the associated VS. The total number J of VMAs relies directly with the order of reflections R. See [8] for a similar adaptation with classical beamforming algorithm.

Applying the BOMP under reverberant conditions relies on this concept With the *Image-Microphone Method* (IMM), the process is performed not only with the real antenna, but also with all the considered VMAs. The implementation of the IMM stills straightforward. The computation of the parameters, like the reflection coefficients β_j or the spherical coordinates $(r^j_{mn}, \Omega^j_{mn})$ of the j^{th} VMA, is analog to those for the ISM.

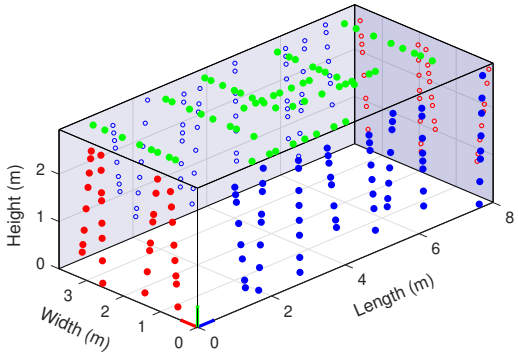


Figure 2: Apparent shape of the array. Only 256 nodes are plotted for visibility

3. EXPERIMENT

The acquisition system reposes on the use of digital MEMS microphones (InvenSense ICS43432). The integration of the conditioning circuitry and A/D conversion directly on the captor highly reduces the whole hardware complexity. Microphones are gathered by eight and plugged on a buffer to form beams. This buffers transmit the pressure datas to a sound card by RJ45 cables. Using USB3 protocole to communicate with computers, the plug-and-play interface allows the management of a huge numbers of sensors synchronously. More informations are provided in [9].

The array takes place in a rectangular room of $8.01m \times 3.75m \times 2.94m$. It comprises 1024 microphones directly flush-mounted to the walls and the ceiling in a pseudo-random distribution. The Figure 2 shows the in-situ proposed array . Only the floor is free of captors for more practicability during measurements. Its deployment induces errors in the position of the MEMS, compared to the predictions. These misalignments can affect the linear dependencies of the propagation model in (1). An acoustical geometric calibration permits to recover the effective geometry of the array. The steering matrix is then adjusted in the inversion in (7). We choose here the robust TOA-based algorithm in [10].

3.1 Source calibration

The asserting of our proposed framework requires to study sources according to the same SH basis of reference. A DP calibration procedure is thus performed before this experiment. The protocol employs a semicircular microphone array of radius $r = 1.19m$, similar to the one described in [11]. The pressure field from the tested source is captured by 104 MEMS microphones regularly spaced on the arc. A turntable rotates the source, to finally reproduce the spherical pressure distribution on a whole sphere. Due to

the experimental setup, we choose to calibrate a controllable electroacoustic source. It brings together two naked loudspeakers, both facing in opposite directions with a spacing between them. Because of the theorical dipolar nature of each, the producing DP should be quadripolar. The corresponding SH vector $\boldsymbol{\alpha}_{ref}$ serves here as a reference. The term $(\cdot)_{expe}$ denotes results from the present experiment. The following correlation factor indicates the quality of the source identification :

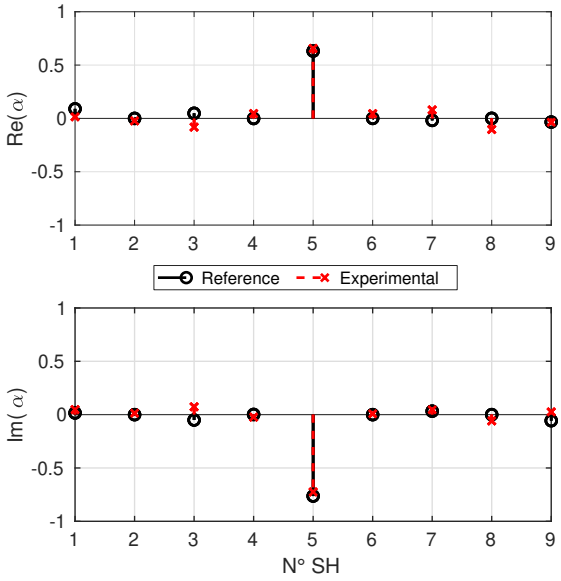
$$\gamma = \frac{< \mathbf{p}_{ref}, \mathbf{p}_{expe} >}{\|\mathbf{p}_{ref}\|_2 \, \|\mathbf{p}_{expe}\|_2} \tag{8}$$

The pressure vectors $\mathbf{p}_{ref}$ and $\mathbf{p}_{expe}$ are both reconstructed with $\boldsymbol{\alpha}_{ref}$ and $\boldsymbol{\alpha}_{expe}$ using (1).

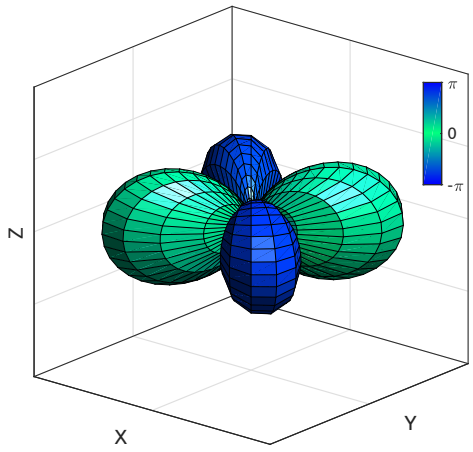
3.2 Results

The source locates now inside the room. The signal used to drive it is a [50Hz-10kHz] logarithmic chirp. The 1024 MEMS acquire synchronously the producing sound field. In addition, a reference microphone is placed inside the domain. Its frequency response and position are known.

The Figure 3a shows the result of the present experiment, using the BOMP coupled with IMM, at frequency $f = 500$Hz. The reflection order is $R = 1$, generating 6 mirrors of the array. The maximum order sets to $L = 2$, describing the DP with 9 SH coefficients. The comparison between $\boldsymbol{\alpha}_{ref}$ and $\boldsymbol{\alpha}_{expe}$ is realized in term of real and imaginary parts. The x-axis represents the linear indexing of the SH basis ($l^2 + l + q + 1$). In both case, the same mode Y_2^{-2} is activated. The expected quadripolar behaviour is then accurately identified, as shown in Figure 3b. The correlation factor gives 96%, significating good agreements with the reference.



(a) SH decomposition of the source



(b) 3D view of the source

Figure 3: Results for the identification of the quadripolar source. The DP estimated in the present experiment $\boldsymbol{\alpha}_{expe}$ (in red) is compared to the reference $\boldsymbol{\alpha}_{ref}$ (in black).

4. CONCLUSION

This paper describes a numerical procedure, validated by an experiment, in order to identify the directivity pattern of complex acoustic sources in real environments. An improvement of the BOMP is proposed for reverberant conditions by considering early reflections to solve the inverse problem. It remains straightforward and easy to implement. In parallel, an acquisition system comprising 1024 synchronous digital MEMS microphones is deploying in a typical rectangular room. The characterization of a quadripolar source is carried out. Results show good agreements with reference measurements and assert the efficiency of the proposed method. However, validations needs to be done for more scenarii, for example with many sources emitting at the same time or showing more complex radiativity. Uncontroled sound sources, like musical instruments, can also be investigated.

5. REFERENCES

[1] M. Noisternig and F. Zotter, "On the decomposition of acoustic source radiation patterns measured with surrounding spherical microphone arrays," *Proc. 37th DAGA*, 2011.

[2] G. Behler, M. Pollow, and M. Vorländer, "Measurements of musical instruments with surrounding spherical arrays," in *Acoustics 2012*, 2012.

[3] T. Ajdler, L. Sbaiz, and M. Vetterli, "The plenacoustic function and its sampling," *IEEE Transactions on Signal Processing*, vol. 54, no. 10, pp. 3790–3804, 2006.

[4] Y. C. Eldar, P. Kuppinger, and H. Bolcskei, "Block-sparse signals: Uncertainty relations and efficient recovery," *IEEE Transactions on Signal Processing*, vol. 58, no. 6, pp. 3042–3054, 2010.

[5] R. Mignot, L. Daudet, and F. Ollivier, "Room reverberation reconstruction: Interpolation of the early part using compressed sensing," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 11, pp. 2301–2312, 2013.

doi:10.25836/sasp.2019.17

[6] H. F. Silverman and W. R. Patterson, "Visualizing the performance of large-aperture microphone arrays," in *1999 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings. ICASSP99 (Cat. No. 99CH36258)*, vol. 2, pp. 969–972, IEEE, 1999.

[7] E. G. Williams, *Fourier acoustics: sound radiation and nearfield acoustical holography*. Elsevier, 1999.

[8] A. Meyer, M. Pelz, and D. Dobler, "Microphone arrays an a wind tunnel environment with a hard reflective floor," in *INTER-NOISE and NOISE-CON Congress and Conference Proceedings*, vol. 258, pp. 3520–3530, Institute of Noise Control Engineering, 2018.

[9] C. Vanwynsberghe, R. Marchiano, F. Ollivier, P. Challande, H. Moingeon, and J. Marchal, "Design and implementation of a multi-octave-band audio camera for realtime diagnosis," *Applied Acoustics*, vol. 89, pp. 281–287, 2015.

[10] C. Vanwynsberghe, P. Challande, F. Ollivier, J. Marchal, and R. Marchiano, "Geometric calibration of very large microphone arrays in mismatched free field," *The Journal of the Acoustical Society of America*, vol. 145, no. 1, pp. 215–227, 2019.

[11] F. Ollivier, A. Peillot, G. Chardon, and L. Daudet, "Acoustic sources joint localization and characterization using compressed sensing," *The Journal of the Acoustical Society of America*, vol. 131, no. 4, pp. 3257–3257, 2012.

doi:10.25836/sasp.2019.17

ONLINE DOA ESTIMATION USING REAL EIGENBEAM ESPRIT WITH PROPAGATION VECTOR MATCHING

Adrian Herzog
International Audio Laboratories Erlangen*
adrian.herzog@audiolabs-erlangen.de

Emanuël A.P. Habets
International Audio Laboratories Erlangen*
emanuel.habets@audiolabs-erlangen.de

ABSTRACT

Eigenbeam ESPRIT (EB-ESPRIT) is a method to estimate multiple directions-of-arrival (DOAs) of sound sources from the spherical harmonics domain (SHD) coefficients of a spherical microphone array recording. Recently, an EB-ESPRIT variant based on three types of recurrence relations of complex spherical harmonics was proposed (DOA-vector EB-ESPRIT). However, due to the signal subspace computation and the joint diagonalization procedure, the computational cost might be too large for many real-time applications. In this work, we propose a computationally more efficient real-valued DOA-vector EB-ESPRIT. The signal subspace is estimated by the deflated projection approximation subspace tracking (PASTd) method. To avoid the joint diagonalization, we propose a subspace propagation-vector matching for the estimation of two DOAs. In the evaluation, we compare the performance of the complex and real DOA-vector EB-ESPRIT with an existing robust B-format DOA estimation method under noisy and reverberant conditions.

1. INTRODUCTION

For parametric time-frequency-domain spatial audio coding, the directions-of-arrival (DOAs) of sound sources have to be estimated from the microphone signals for each time-frame and frequency band. Accurately estimating these parameters in real-time is a challenging task.

For directional audio coding (DirAC) [1], one DOA has to be estimated per time-frame and frequency band, which can be done with low computational cost using the pseudointensity vector (PIV) [2]. The accuracy of the PIV is, however, quite limited compared to subspace-based DOA estimators [3]. For high angular resolution plane-wave expansion (HARPEX) [4], two DOAs are estimated from a B-format signal per time-frame and frequency band using properties of plane-wave propagation vectors. In [5], a robust version of this DOA estimator is proposed which computes a signal subspace prior to the DOA estimation.

Eigenbeam ESPRIT [6] uses recurrence relations of spherical harmonics to estimate multiple DOAs from a spherical harmonics domain signal. In [7–10], robust and unambiguous EB-ESPRIT variants have been developed. In [10], the authors proposed the DOA-vector EB-ESPRIT which can accurately estimate the DOAs. However, the computational complexity of the DOA-vector EB-ESPRIT is high due to the signal subspace estimation and the joint diagonalization procedure. Hence, a computationally more efficient version of the DOA-vector EB-ESPRIT is needed.

In this work, we propose a new DOA-vector EB-ESPRIT based on real spherical harmonics recurrence relations and the computationally efficient deflated projection approximation subspace tracking (PASTd) algorithm [11]. If only one DOA has to be estimated per time-frame and frequency band, the EB-ESPRIT equations can be simplified. For estimating two DOAs, we propose to first estimate the plane-wave propagation vectors using properties of the real spherical harmonics and then apply the simplified EB-ESPRIT equations for one DOA to both estimated propagation vectors, thereby, avoiding the joint diagonalization. This procedure is referred to as subspace propagation-vector matching in the remainder of this work.

In Sec. 2, spherical harmonic domain signals are introduced. In Sec. 3, the real DOA-vector EB-ESPRIT is derived. In Sec. 4, simplifications for an efficient online implementation are discussed. In Sec. 5, the proposed method is evaluated and compared to [10] and [5].

2. SPHERICAL HARMONICS DOMAIN

Let $p(k; r, \Omega)$ denote the sound pressure field on the surface of a spherical microphone array (SMA) with radius r, where k denotes the wavenumber and $\Omega = (\theta, \phi)$ the angular position on the sphere specified by the elevation $\theta \in [0, \pi]$ and azimuth $\phi \in [-\pi, \pi)$ angles. The pressure field can be expanded using the spherical harmonics expansion [12, 13] as

$$p(k; r, \Omega) = \sum_{l=0}^{\infty} \sum_{m=-l}^{l} b_l(kr) P_{lm}(k) Y_{lm}(\Omega) \,, \quad (1)$$

where l and m denote the *order* and *mode*, respectively. The radial dependencies $b_l(kr)$ are denoted as *mode strengths* which depend on the SMA properties only,

*A joint institution of the Friedrich Alexander University Erlangen-Nürnberg (FAU) and Fraunhofer IIS, Germany.

© Adrian Herzog, Emanuël A.P. Habets. Licensed under a Creative Commons Attribution 4.0 International License (CC BY 4.0). **Attribution:** Adrian Herzog, Emanuël A.P. Habets. "Online DOA Estimation Using Real Eigenbeam ESPRIT with Propagation Vector Matching", 1st EAA Spatial Audio Signal Processing Symposium, Paris, France, 2019.

doi:10.25836/sasp.2019.18

$P_{lm}(k)$ are the spherical harmonic domain (SHD) coefficients and $Y_{lm}(\Omega)$ the complex spherical harmonic functions [12, 13].

The real spherical harmonics $R_{lm}(\Omega)$ are related to $Y_{lm}(\Omega)$ as follows [13]:

$$R_{lm}(\Omega) = \begin{cases} \frac{i}{\sqrt{2}}\left(Y_{lm}(\Omega) - (-1)^m Y_{l(-m)}(\Omega)\right) & m < 0 \\ Y_{l0}(\Omega) & m = 0 \\ \frac{1}{\sqrt{2}}\left(Y_{l(-m)}(\Omega) + (-1)^m Y_{lm}(\Omega)\right) & m > 0 \end{cases} \tag{2}$$

with $i^2 = -1$. Using the vector notation

$$\begin{aligned} \mathbf{y}(\Omega) &:= [Y_{00}, Y_{1-1}, Y_{10}, ..., Y_{LL}]^T(\Omega) \text{ and} \\ \mathbf{r}(\Omega) &:= [R_{00}, R_{1-1}, R_{10}, ..., R_{LL}]^T(\Omega)\,, \end{aligned} \tag{3}$$

where $(\cdot)^T$ denotes the transpose and L the maximum SHD order considered, we can write

$$\mathbf{r}(\Omega) = \mathbf{U}_L \mathbf{y}(\Omega) = \mathbf{U}_L^* \mathbf{y}^*(\Omega) \tag{4}$$

with the unitary $(L+1)^2 \times (L+1)^2$ matrix $\mathbf{U}_L$ and $(\cdot)^*$ denoting the complex conjugate. Let us define:

$$\begin{aligned} \mathbf{p}(k) &:= [P_{00}, P_{1-1}, P_{10}, P_{11}, ..., P_{LL}]^T(k) \text{ and} \\ \mathbf{B}(kr) &:= \text{diag}\left\{[b_0, b_1, b_1, b_1, ..., b_L]\right\}(kr)\,, \end{aligned} \tag{5}$$

where in $\mathbf{B}(kr)$, each mode strength $b_l(kr)$ appears $2l+1$ times on the diagonal. The complex and real *spherical harmonics transforms* can then be defined as follows:

$$\begin{aligned} \mathbf{p}(k) &= \mathbf{B}^{-1}(kr) \int_{\mathcal{S}^2} p(k; r, \Omega) \mathbf{y}^*(\Omega)\, \mathrm{d}\Omega \\ \mathbf{p}^{\mathfrak{R}}(k) &= \mathbf{B}^{-1}(kr) \int_{\mathcal{S}^2} p(k; r, \Omega) \mathbf{r}(\Omega)\, \mathrm{d}\Omega\,. \end{aligned} \tag{6}$$

In practice, the mode-strength compensation $\mathbf{B}^{-1}(kr)$ has to be regularized and the integral has to be approximated by a weighted sum over the microphone directions $\Omega_1, ..., \Omega_P$ of the SMA. Moreover, the sound pressure at these directions has to be replaced with the respective microphone signals $X_1, ..., X_P$. This yields the discrete spherical harmonics transform:

$$\mathbf{x}(k) = \mathbf{B}_{\text{reg}}^{-1}(kr) \sum_{p=1}^{P} \mathfrak{q}_p X_p(k) \mathbf{y}^*(\Omega_p) \tag{7}$$

and analogously for the real spherical harmonics transform, where $\mathbf{B}_{\text{reg}}^{-1}$ denotes the regularized inverse of $\mathbf{B}$ and $\mathfrak{q}_1, ..., \mathfrak{q}_P$ the sampling weights, which depend on the microphone distribution of the SMA [12, 13]. For uniform spatial sampling, one yields $\mathfrak{q}_p = \frac{4\pi}{P}$ [14].

In the following sections, we assume that all signals have been transformed to the short-time Fourier transform (STFT) domain, where time and frequency indices are omitted for brevity.

3. DOA-VECTOR EB-ESPRIT

3.1 Complex DOA-Vector EB-ESPRIT

Let $\mathbf{x} = [X_{00}, X_{1-1}, ..., X_{LL}]$ denote the mode strength compensated complex SHD coefficients of a SMA recording including J plane-wave sources and additive noise, which are assumed to be mutually uncorrelated. The power spectral density (PSD) matrix of $\mathbf{x}$ is defined as $\mathbf{\Phi}_{\mathbf{x}} := E\{\mathbf{x}\mathbf{x}^H\}$, where $(\cdot)^H$ denotes the conjugate transpose and $E\{\cdot\}$ the statistical expectation. It can be shown that the plane-wave propagation vectors are proportional to $\mathbf{y}^*(\Omega_j)$ for $j = 1, ..., J$, where Ω_j denotes the DOA of source j [12]. The signal subspace span$\{\mathbf{y}^*(\Omega_1), ..., \mathbf{y}^*(\Omega_J)\}$ can be estimated from the eigenvectors $[\mathbf{u}_1, ..., \mathbf{u}_J] =: \mathbf{U}_s$ corresponding to the J largest eigenvalues using the following relation [6]:

$$\mathbf{U}_s = [\mathbf{y}^*(\Omega_1), ..., \mathbf{y}^*(\Omega_J)]\, \mathbf{T}\,, \tag{8}$$

where $\mathbf{T}$ is an invertible matrix of size $J \times J$.

EB-ESPRIT uses (8) and recurrence relations of spherical harmonics to estimate the plane-wave DOAs $\Omega_1, ..., \Omega_J$ from the signal subspace eigenmatrix $\mathbf{U}_s$. The DOA-vector EB-ESPRIT [10] estimates the DOA-vectors $\mathbf{n}(\Omega_1), ..., \mathbf{n}(\Omega_J)$, defined as

$$\mathbf{n}(\Omega) = \begin{bmatrix} n_x(\Omega) \\ n_y(\Omega) \\ n_z(\Omega) \end{bmatrix} = \begin{bmatrix} \sin(\theta)\cos(\phi) \\ \sin(\theta)\sin(\phi) \\ \cos(\theta) \end{bmatrix} \tag{9}$$

using the following three recurrence relations:

$$n_a(\Omega)\mathbf{D}_0 \mathbf{y}^*(\Omega) = \mathbf{D}_a \mathbf{y}^*(\Omega) \text{ for } a = x, y, z \tag{10}$$

with

$$\mathbf{D}_x := \tfrac{1}{2}(\mathbf{D}_- + \mathbf{D}_+) \quad , \quad \mathbf{D}_y := \tfrac{1}{2i}(\mathbf{D}_- - \mathbf{D}_+)\,, \tag{11}$$

and

$$\begin{aligned} [\mathbf{D}_z(\cdot)]_{lm} &= \sqrt{(l-m)(l+m)/(2l-1)(2l+1)}\, [(\cdot)]_{(l-1)m} \\ &\quad + \sqrt{(l+1-m)(l+1+m)/(2l+1)(2l+3)}\, [(\cdot)]_{(l+1)m} \\ [\mathbf{D}_\pm(\cdot)]_{lm} &= \pm\sqrt{(l-1\mp m)(l\mp m)/(2l-1)(2l+1)}\, [(\cdot)]_{(l-1)(m\pm 1)} \\ &\quad \mp\sqrt{(l+1\pm m)(l+2\pm m)/(2l+1)(2l+3)}\, [(\cdot)]_{(l+1)(m\pm 1)} \\ \mathbf{D}_0(\cdot) &= [[(\cdot)]_{00}, ..., [(\cdot)]_{(L-1)(L-1)}]^T \end{aligned} \tag{12}$$

for $l = 0, ..., L-1$ and $m = -l, ..., l$, where $[(\cdot)]_{lm} := 0$ if $|m| > l$. Using relation (8), we get from (10)

$$(\mathbf{D}_0 \mathbf{U}_s)\mathbf{\Psi}_a = \mathbf{D}_a \mathbf{U}_s \tag{13}$$

with $\mathbf{\Psi}_a = \mathbf{T}^{-1}\text{diag}\{n_a(\Omega_1), ..., n_a(\Omega_J)\}\mathbf{T}$ for $a = x, y, z$. The matrices $\mathbf{\Psi}_x$, $\mathbf{\Psi}_y$ and $\mathbf{\Psi}_z$ can be estimated in the least-squares sense by:

$$\hat{\mathbf{\Psi}}_a = (\mathbf{D}_0 \mathbf{U}_s)^+ \mathbf{D}_a \mathbf{U}_s\,, \tag{14}$$

where $(\cdot)^+$ denotes the pseudo-inverse. The DOA-vectors $\mathbf{n}(\Omega_1), ..., \mathbf{n}(\Omega_J)$ can then be estimated by jointly diagonalizing $\hat{\mathbf{\Psi}}_x$, $\hat{\mathbf{\Psi}}_y$ and $\hat{\mathbf{\Psi}}_z$ as follows [10]:

$$[\hat{\mathbf{n}}(\Omega_1), ..., \hat{\mathbf{n}}(\Omega_J)] = \text{Re}\left\{[\boldsymbol{\lambda}_x, \boldsymbol{\lambda}_y, \boldsymbol{\lambda}_z]^T\right\}\,, \tag{15}$$

where $\boldsymbol{\lambda}_x$, $\boldsymbol{\lambda}_y$ and $\boldsymbol{\lambda}_z$ denote the eigenvalue vectors of $\hat{\mathbf{\Psi}}_x$, $\hat{\mathbf{\Psi}}_y$ and $\hat{\mathbf{\Psi}}_z$, respectively, derived using a joint diagonalization method.

doi:10.25836/sasp.2019.18

3.2 Real DOA-Vector EB-ESPRIT

Let $\mathbf{x}^{\mathfrak{R}}$ denote the real SHD signal corresponding to $\mathbf{x}$. Using $\mathbf{x}^{\mathfrak{R}} = \mathbf{U}_L^* \mathbf{x}$ and (4), one can show that the plane-wave propagation vectors are proportional to $\mathbf{r}(\Omega_j)$ for $j = 1, ..., J$. Note that, although the propagation vectors are real valued, the signal vectors are still complex valued due to the complex STFT. Analogously to [5], we split $\mathbf{x}^{\mathfrak{R}}$ into real and imaginary parts and construct a real valued PSD matrix thereof:

$$\mathbf{X} := [\mathrm{Re}\{\mathbf{x}^{\mathfrak{R}}\}, \mathrm{Im}\{\mathbf{x}^{\mathfrak{R}}\}] \qquad \mathbf{\Phi}_{\mathbf{X}}^{\mathfrak{R}} := E\{\mathbf{X}\mathbf{X}^H\} \,. \quad (16)$$

One can show that $\mathbf{\Phi}_{\mathbf{X}}^{\mathfrak{R}}$ coincides with the PSD matrix of $\mathbf{x}^{\mathfrak{R}}$, if the noise PSD matrix is real valued. Theoretically, this is the case for diffuse noise and microphone self-noise.

The real signal subspace is constructed from the real eigenvectors $[\mathbf{o}_1, ..., \mathbf{o}_J] =: \mathbf{O}_s$ corresponding to the J largest eigenvalues of $\mathbf{\Phi}_{\mathbf{X}}^{\mathfrak{R}}$, which is related to the plane-wave propagation vectors as follows:

$$\mathbf{O}_s = [\mathbf{r}(\Omega_1), ..., \mathbf{r}(\Omega_J)]\, \mathbf{T}^{\mathfrak{R}} \,, \quad (17)$$

where $\mathbf{T}^{\mathfrak{R}}$ is a real-valued invertible $J \times J$ matrix. Using relation (4), the recurrence relations (10) can be formulated for real spherical harmonics:

$$\begin{aligned} n_a(\Omega)\mathbf{D}_0\mathbf{y}^*(\Omega) &= \mathbf{D}_a\mathbf{y}^*(\Omega) \\ n_a(\Omega)\mathbf{D}_0\mathbf{U}_L^T\mathbf{r}(\Omega) &= \mathbf{D}_a\mathbf{U}_L^T\mathbf{r}(\Omega) \\ n_a(\Omega)\mathbf{U}_{L-1}^T\mathbf{D}_0\mathbf{r}(\Omega) &= \mathbf{D}_a\mathbf{U}_L^T\mathbf{r}(\Omega) \\ n_a(\Omega)\mathbf{D}_0\mathbf{r}(\Omega) &= \mathbf{U}_{L-1}^*\mathbf{D}_a\mathbf{U}_L^T\mathbf{r}(\Omega) =: \mathbf{D}_a^{\mathfrak{R}}\mathbf{r}(\Omega) \,, \end{aligned} \quad (18)$$

where we defined $\mathbf{D}_a^{\mathfrak{R}} := \mathbf{U}_{L-1}^*\mathbf{D}_a\mathbf{U}_L^T$ which must be real valued for $a = x, y, z$. Using (17), the EB-ESPRIT equations become:

$$(\mathbf{D}_0\mathbf{O}_s)\mathbf{\Psi}_a^{\mathfrak{R}} = \mathbf{D}_a^{\mathfrak{R}}\mathbf{O}_s \,. \quad (19)$$

The DOA-vectors can then be estimated analogously to (14)-(15). In contrast to the complex DOA-vector EB-ESPRIT, the matrices involved are real valued reducing the computational complexity.

4. ONLINE IMPLEMENTATION

In principle, one can use the real or complex DOA-vector EB-ESPRIT to estimate the source DOAs per time-frequency bin in an online manner. However, due to the eigendecompostion of the PSD matrix and the joint diagonalization procedure, the computational cost might be too large for many real-time applications. Therefore, we propose various modifications in the following sections.

4.1 Recursive Subspace Tracking

Let n and k denote the time-frame and frequency indices, respectively. The PSD matrix of $\mathbf{x}(n,k)$ or $\mathbf{X}(n,k)$ can be estimated recursively as follows:

$$\begin{aligned} &\hat{\mathbf{\Phi}}_{\mathbf{x}}(n,k) = \beta\hat{\mathbf{\Phi}}_{\mathbf{x}}(n-1,k) + (1-\beta)\mathbf{x}(n,k)\mathbf{x}^H(n,k) \text{ or} \\ &\hat{\mathbf{\Phi}}_{\mathbf{X}}^{\mathfrak{R}}(n,k) = \beta\hat{\mathbf{\Phi}}_{\mathbf{X}}^{\mathfrak{R}}(n-1,k) + (1-\beta)\mathbf{X}(n,k)\mathbf{X}^H(n,k) \,, \end{aligned} \quad (20)$$

Algorithm 1: PASTd for real DOA-vector EB-ESPRIT

```
X = X(n,k);
for j = 1, ..., J do
    z^T = o_j^T(n-1,k) X;
    λ_j(n,k) = β λ_j(n-1,k) + ||z||^2;
    E = X - o_j(n-1,k) z^T;
    o_j(n,k) = o_j(n-1,k) + E z^* / λ_j(n,k);
    X = X - o_j(n,k) z^T;
```

where $\beta \in [0, 1)$ is a forgetting factor. The signal subspace can be constructed by performing a singular value decomposition (SVD) to $\hat{\mathbf{\Phi}}_{\mathbf{x}}(n,k)$ or $\hat{\mathbf{\Phi}}_{\mathbf{X}}^{\mathfrak{R}}(n,k)$ and then selecting the eigenvectors corresponding to the J largest eigenvalues. However, this involves a SVD of a $(L+1)^2 \times (L+1)^2$ matrix per time-frequency bin.

A computationally more efficient method to recursively estimate the signal subspace is the deflated projection approximation subspace tracking (PASTd) algorithm [11]. The PASTd algorithm for real SHD signals is summarized in Alg. 1. Note, that $\mathbf{X}(n,k)$ is a $(L+1)^2 \times 2$ matrix and thus the PASTd from [11] has been adjusted accordingly.

Additionally, we othonormalize the estimated eigenvectors using a QR decomposition which can be implemented with low computational cost using e.g. the modified Gram-Schmidt algorithm [15].

4.2 Simplifications for $J = 1$

For $J = 1$, the signal subspace is one dimensional and the matrices $\hat{\mathbf{\Psi}}_x$, $\hat{\mathbf{\Psi}}_y$ and $\hat{\mathbf{\Psi}}_z$ become scalars $\hat{\Psi}_x$, $\hat{\Psi}_y$ and $\hat{\Psi}_z$. Therefore, no joint diagonalization is necessary. The DOA-vector can be estimated directly via:

$$\begin{aligned} \hat{\mathbf{n}}(\Omega_1) &= \mathrm{Re}\{[\hat{\Psi}_x, \hat{\Psi}_y, \hat{\Psi}_z]^T\} \\ &= \|\mathbf{D}_0\mathbf{u}_1\|^{-2}\mathrm{Re}\left\{[\mathbf{D}_x\mathbf{u}_1, \mathbf{D}_y\mathbf{u}_1, \mathbf{D}_z\mathbf{u}_1]^T\right\} \,, \end{aligned} \quad (21)$$

where $\mathbf{u}_1$ is the dominant eigenvector of $\mathbf{\Phi}_{\mathbf{x}}$. The factor $\|\mathbf{D}_0\mathbf{u}_1\|^{-2}$ can be replaced by a normalization to ensure $\|\hat{\mathbf{n}}(\Omega_1)\| = 1$. These simplifications can be made for the real DOA-vector EB-ESPRIT analogously.

4.3 Subspace Propagation-Vector Matching

In [4, 5], general properties of plane-wave propagation vectors are employed to estimate two DOAs per time-frequency bin from a B-format signal. In this section, we develop a similar method to estimate the mixing matrix $\mathbf{C} := (\mathbf{T}^{\mathfrak{R}})^{-1}$ for the real DOA-vector EB-ESPRIT with two sources ($J = 2$). From (17) we get:

$$[\mathbf{r}(\Omega_1), \mathbf{r}(\Omega_2)] = \mathbf{O}_s\mathbf{C} \,. \quad (22)$$

We use the following properties of real spherical harmonics:

$$R_{00}(\Omega) = \frac{1}{\sqrt{4\pi}} \qquad \|\mathbf{r}_1(\Omega)\|^2 = \frac{1}{\pi} \,, \quad (23)$$

where $\mathbf{r}_1(\Omega)$ is the vector of real spherical harmonics up to order 1. Note, that the conditions (23) are sufficient to ensure that $\mathbf{r}_1(\Omega)$ is a real spherical harmonic vector. To ensure that $\mathbf{r}(\Omega)$ describes a real spherical harmonic vector,

we would need more conditions. Let us denote with $\mathbf{O}_s^{(1)}$ the coefficients of $\mathbf{O}_s$ up to first order and with $\mathbf{Q}^{(1)}\mathbf{R}^{(1)}$ the QR decomposition thereof. Using (22), we find

$$[\mathbf{r}_1(\Omega_1), \mathbf{r}_1(\Omega_2)] = \mathbf{Q}^{(1)}\tilde{\mathbf{C}} = \mathbf{Q}^{(1)} \begin{bmatrix} \tilde{c}_{11} & \tilde{c}_{12} \\ \tilde{c}_{21} & \tilde{c}_{22} \end{bmatrix} , \quad (24)$$

where $\tilde{\mathbf{C}} = \mathbf{R}^{(1)}\mathbf{C}$. Inserting (24) into the conditions (23) and using the orthonormal property of $\mathbf{Q}^{(1)}$ yields:

$$\mathbf{q}^T\tilde{\mathbf{c}}_j = \frac{1}{\sqrt{4\pi}} \qquad \text{and} \qquad \|\tilde{\mathbf{c}}_j\|^2 = \frac{1}{\pi} \quad (25)$$

for $j = 1, 2$, where we defined $\mathbf{q}^T := [Q^{(1)}_{00,1}, Q^{(1)}_{00,2}]$ and $\tilde{\mathbf{c}}_j := [\tilde{c}_{1j}, \tilde{c}_{2j}]^T$. Using the second condition, one can write $\tilde{\mathbf{c}}_j$ in the following form:

$$\tilde{\mathbf{c}}_j = \tfrac{1}{\sqrt{\pi}} \left[\cos(\varphi_j), \sin(\varphi_j)\right]^T . \quad (26)$$

Writing $\mathbf{q}$ in terms of magnitude q and phase φ_q:

$$\mathbf{q} = q \left[\cos(\varphi_q), \sin(\varphi_q)\right]^T \quad (27)$$

and inserting (26) into the first condition of (25) one can derive

$$\frac{1}{\sqrt{4\pi}} = \mathbf{q}^T\tilde{\mathbf{c}}_j = \frac{q}{\sqrt{\pi}}\cos(\varphi_q - \varphi_j)$$
$$\Rightarrow \varphi_j = \varphi_q \mp \operatorname{acos}\left(\frac{1}{2q}\right) . \quad (28)$$

We, therefore, get two solutions which can be assigned to φ_1 and φ_2. The results are real valued if $q \geq \frac{1}{2}$. Otherwise, we estimate one DOA only from the dominant eigenvector $\mathbf{o}_1$. The mixing matrix $\mathbf{C}$ can be obtained from φ_1 and φ_2 as follows:

$$\mathbf{C} = (\mathbf{R}^{(1)})^{-1}\tilde{\mathbf{C}} = \frac{1}{\sqrt{\pi}}(\mathbf{R}^{(1)})^{-1} \begin{bmatrix} \cos(\varphi_1) & \cos(\varphi_2) \\ \sin(\varphi_1) & \sin(\varphi_2) \end{bmatrix} . \quad (29)$$

The full higher-order plane-wave propagation vectors $\mathbf{r}(\Omega_1)$ and $\mathbf{r}(\Omega_2)$ can then be estimated by applying $\mathbf{C}$ to the full signal subspace $\mathbf{O}_s$, i.e.,

$$[\hat{\mathbf{r}}(\Omega_1), \hat{\mathbf{r}}(\Omega_2)] = \mathbf{O}_s\mathbf{C} . \quad (30)$$

Finally, the simplified DOA-vector EB-ESPRIT equations discussed in Sec. 4.2 can be used to estimate the DOA-vectors from $\hat{\mathbf{r}}(\Omega_1)$ and $\hat{\mathbf{r}}(\Omega_2)$ separately. Hence, the joint diagonalization is avoided by estimating the source propagation vectors before the EB-ESPRIT equations are used.

As $\mathbf{C}$ is estimated using zero- and first-order coefficients of $\mathbf{O}_s$ only, $\mathbf{O}_s\mathbf{C}$ is not necessarily close to a set of plane-wave propagation vectors. Therefore, we perform a consistency check between the DOA-vectors $\hat{\mathbf{n}}(\Omega_j)$ estimated with the EB-ESPRIT equations and $\hat{\mathbf{n}}_{\text{fo}}(\Omega_j)$ derived from the first order coefficients as follows:

$$\hat{\mathbf{n}}_{\text{fo}}(\Omega_j) := \sqrt{\tfrac{4\pi}{3}}[\hat{R}_{11}(\Omega_j), \hat{R}_{1-1}(\Omega_j), \hat{R}_{10}(\Omega_j)]^T \quad (31)$$

for $J = 1, 2$. If their angular distance is $\geq \Delta\varphi$, we use $\hat{\mathbf{n}}_{\text{fo}}(\Omega)$ instead of $\hat{\mathbf{n}}(\Omega_j)$ for the DOA-vector estimate. The proposed method can be summarized as follows:

1. Update real signal subspace matrix (Sec. 4.1)
2. Estimate plane-wave propagation vectors using subspace propagation-vector matching (Sec. 4.3)
3. Estimate DOA-vectors with simplified $J = 1$ real DOA-vector EB-ESPRIT (Sec. 4.2)
4. Consistency check of DOA-vectors with first order coefficients of estimated propagation vectors (Sec. 4.3, last paragraph)

5. EVALUATION

5.1 Setup

For the evaluation, third order ($L = 3$) spherical harmonic domain signals with one or two plane-wave sources, reverberation and diffuse stationary noise were simulated. The plane-wave source signals consisted of male and female English speech signals of 3.8 seconds length and sampled at 16 kHz, taken from [16]. The source DOAs were randomly and uniquely selected from a set of 48 uniformly distributed directions. For the dual source scenario, the 48 directions were divided into two sectors, from which the DOAs are selected.

For the non-reverberant scenarios, the plane-wave sources were transformed to the STFT domain. Each time-frequency bin was then multiplied with the plane-wave propagation vector $\mathbf{y}^*(\Omega)$ at the corresponding DOA Ω. For the reverberant scenarios, microphone signals of a rigid spherical microphone array with 32 microphones and 7 cm radius placed at $[4.103\text{ m}, 3.471\text{ m}, 2.912\text{ m}]$ in a $8 \times 7 \times 6$ m^3 shoebox room were simulated using [17]. The sources were placed at a distance of 2 m from the virtual microphone array. Reverberation times (T_{60}) of 0.3 and 0.6 seconds were used. The microphone signals were then transformed to the STFT domain. Finally, the SHD coefficients are derived using the discrete spherical harmonics transform (7) with uniform sampling weights and $\mathbf{B}_{\text{reg}}^{-1} = (\mathbf{B}^H\mathbf{B} + \lambda\mathbf{I})^{-1}\mathbf{B}^H$ with $\lambda = 10^{-6}$.

For all cases, diffuse stationary white noise with signal plus reverberation-to-noise ratio SNR $= 6$ dB was added. The desired source variance, which is needed to determine the variance of the noise, was computed as the mean energy of the noiseless signal, excluding time-frames with energies less than 1% of the maximum frame energy.

For the STFTs, a frame-length of 128 samples (8 ms), 50% overlap, a square-root-Hann window and a discrete Fourier transform size of 256 was chosen.

For the recursive signal-subspace estimation, $\beta = 0.9$ ($\hat{=}$ 38 ms time-constant) was chosen. An angular distance $\Delta\varphi = 0.4\pi$ was chosen for the consistency check of the subspace propagation-vector matching. The DOAs were estimated within the frequency range $[100, 2340]$ Hz, were the lower bound has been chosen to reduce the effect of the regularized mode-strength compensation and the upper bound to avoid spatial aliasing.

For the performance evaluation we computed angular

doi:10.25836/sasp.2019.18

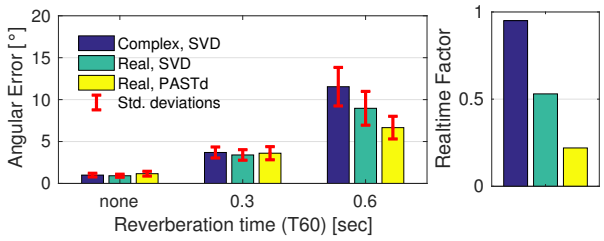


Figure 1. Mean angular estimation errors and mean realtime factors for single source scenario ($J = 1$)

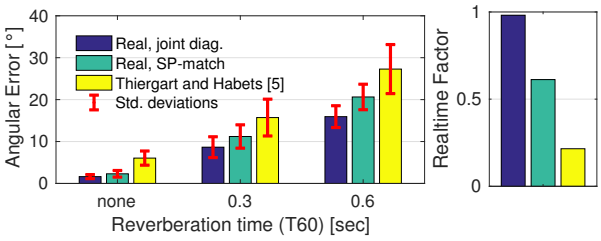


Figure 2. Mean angular estimation errors and mean realtime factors for dual sources scenario ($J = 2$)

estimation errors

$$\Delta\Omega_j(n,k) = \min_{j'} \left\{ \text{acos}\left(\mathbf{n}(\Omega_j)^T \mathbf{n}(\hat{\Omega}_{j'}(n,k)) \right) \right\} , \tag{32}$$

where Ω_j is the true DOA of source j and $\hat{\Omega}_{j'}(n,k)$ the j'-th estimated DOA at time-frequency bin (n,k). The mean angular estimation errors within the regions where the sources are active are defined as

$$\overline{\Delta\Omega_j} = \sum_{n,k} w_j(n,k)\Delta\Omega_j(n,k) , \tag{33}$$

where $w_j(n,k)$ is one if the narrowband frame energy of source j, including reverberation, at bin (n,k) is greater than -30 dB w.r.t. the maximum energy and zero otherwise. To evaluate the computational complexity we computed realtime factors $= \frac{\text{Computation time}}{\text{Signal length}}$. The DOA estimators were implemented with MATLAB [18] in double precision and executed on a computer with a 3.40 GHz CPU.

5.2 Single source scenario

For the single source scenario, 20 SHD signals with one plane-wave source were generated. The source signals consisted of 10 female and 10 male English speech signals.

In Fig. 1, the mean angular estimation errors and realtime factors for the complex DOA-vector EB-ESPRIT with SVD-based subspace estimation and the real DOA-vector EB-ESPRIT with SVD-based subspace estimation or PASTd are shown for different reverberation times. The angular errors have been averaged over the 20 experiments and the corresponding standard deviations are represented with red errorbars.

One can see that the real-valued formulation and the PASTd reduce the computational cost by $\sim 75\%$ without a significant loss of DOA estimation accuracy.

For $T_{60} = 0.6$ seconds, the proposed real DOA-vector EB-ESPRIT yields less estimation errors than the complex DOA-vector EB-ESPRIT. Recall that $\mathbf{\Phi}_{\mathbf{X}}^{\mathfrak{R}} = \text{Re}\{\mathbf{\Phi}_{\mathbf{x}^{\mathfrak{R}}}\}$. The PSD matrices of the direct-path plane-wave sources and the diffuse noise are real-valued in the real SHD. Only the early and late reflections may contribute to $\text{Im}\{\mathbf{\Phi}_{\mathbf{x}^{\mathfrak{R}}}\}$. Therefore, it is plausible that it is more robust to use $\mathbf{\Phi}_{\mathbf{X}}^{\mathfrak{R}}$ instead of $\mathbf{\Phi}_{\mathbf{x}^{\mathfrak{R}}}$ for the subspace estimation under reverberant conditions. Using the PASTd instead of the SVD further improves the estimation accuracy for $T_{60} = 0.6$ seconds, which can be explained by the fact that PASTd can be more robust than the SVD for low signal to noise/reverberation ratios [11].

5.3 Dual sources scenario

For the dual sources scenario, the real DOA vector EB-ESPRIT with PASTd and joint diagonalization (joint diag.) or subspace propagation-vector matching (SP-match) were compared with a robust B-format DOA estimator [5]. Ten different SHD signals with one female and one male plane-wave source were generated.

The mean angular errors and realtime factors are shown in Fig. 1. The results are averaged over the 10 configurations and the two DOAs. The respective standard deviations are represented with red errorbars. The EB-ESPRIT-based methods yield lower estimation errors compared to the robust B-format method, which is expected as the latter does not incorporate higher-order SHD coefficients. The real DOA-vector EB-ESPRIT with SP-matching is less accurate than the joint diagonalization based method. However, the computational cost is reduced by $\sim 40\%$.

So far only mean angular estimation errors have been analysed. In what follows, we analyse a dual source scenario with DOAs $\Omega_1 = [90°, 60°]$, $\Omega_2 = [130°, -80°]$ and $T_{60} = 0.3$ seconds in more detail.

In Fig. 3, the spectrogram of the source signals (a), the source activity per time-frequency bin (b) and angular estimation errors for both source DOAs (c-h) are shown. One can see that the EB-ESPRIT-based methods yield less estimation errors than the robust B-format method, except at time-frequency bins where the number of active sources changes from one to two are two to one. For the robust B-format method, these regions are less critical, however, the overall angular estimation errors are larger.

In Fig. 4, distributions of the estimated azimuth and elevation angles are shown for the three methods. One can see that, for the EB-ESPRIT-based methods, the estimated azimuth and elevation are mostly concentrated around the true DOAs ($\Omega_1 = [90°, 60°]$, $\Omega_2 = [130°, -80°]$), while for the robust B-format method, the estimates are more scattered across the angular space.

6. CONCLUSION

We proposed the real DOA-vector EB-ESPRIT which reduces the computational complexity of the DOA-vector EB-ESPRIT [10] by working with real-valued quantities and by efficiently estimating the signal subspace using the PASTd algorithm [11]. To futher reduce the computational complexity, we replaced the joint diagonalization with a

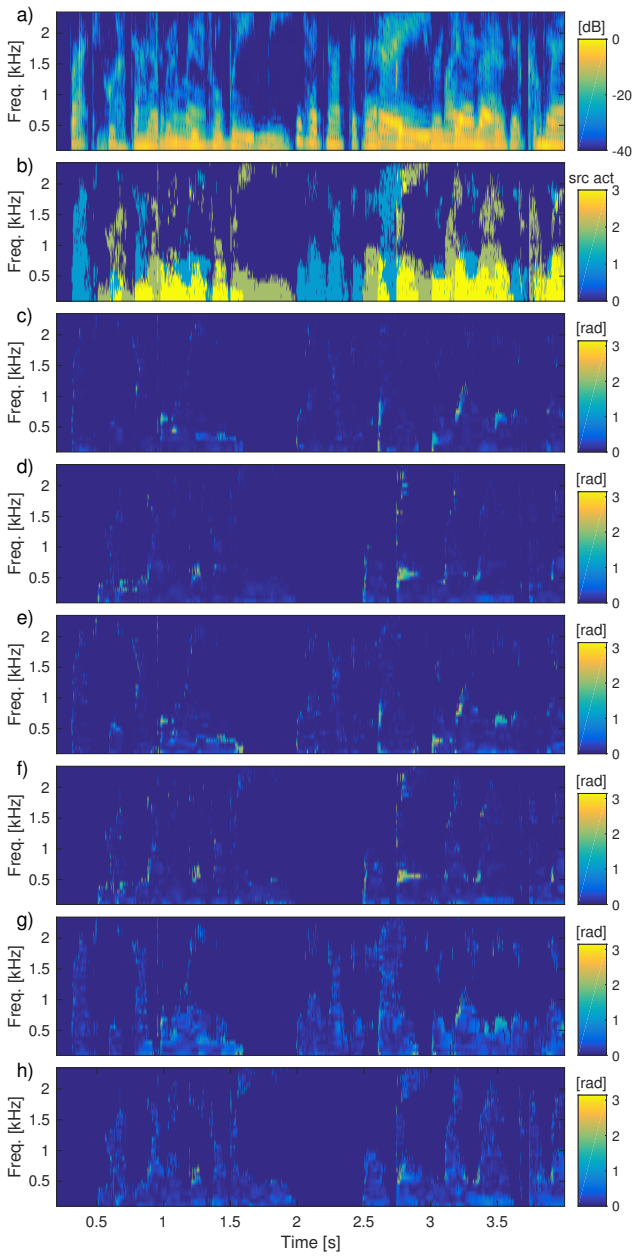


Figure 3. a) Spectrogram of source signals (sum), b) active time-frequency bins with 0: no source active, 1: source 1 active, 2: source 2 active, 3: both sources active, c) - h): Angular estimation errors at active bins for Ω_1 and Ω_2, c) and d) Real, joint diag., e) and f) Real, SP-match, g) and h) Thiergart and Habets [5]

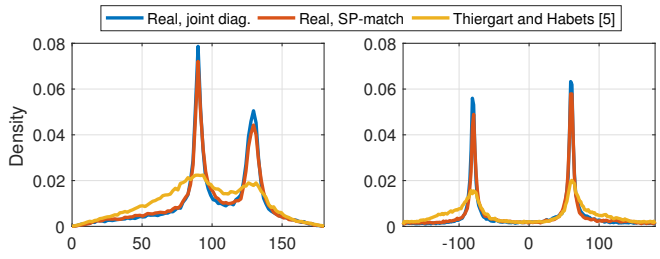


Figure 4. Distributions of elevation and azimuth estimates

subspace propagation-vector matching method for estimating two DOAs. In the evaluation, we showed that the computational cost of the proposed method is significantly reduced compared to the complex DOA-vector EB-ESPRIT and that the method can estimate the source DOAs more accuratly compared to [5].

7. REFERENCES

[1] V. Pulkki, "Spatial sound reproduction with directional audio coding," *Journal Audio Eng. Soc.*, vol. 55, pp. 503–516, June 2007.

[2] D. P. Jarrett, E. A. P. Habets, and P. A. Naylor, "3D source localization in the spherical harmonic domain using a pseudointensity vector," in *Proc. European Signal Processing Conf. (EUSIPCO)*, (Aalborg, Denmark), pp. 442–446, Aug. 2010.

[3] A. Herzog and E. A. P. Habets, "On the relation between DOA-vector eigenenbeam ESPRIT and subspace-pseudointensity-vector," in *Proc. European Signal Processing Conf. (EUSIPCO)*, (A Coruña, Spain), Sept. 2019.

[4] S. Berge and N. Barrett, "High angular resolution planewave expansion," in *2nd Intl. Symp. on Ambisonics and Spherical Acoustics*, (Paris, France), May 2010.

[5] O. Thiergart and E. A. P. Habets, "Robust direction-of-arrival estimation of two simultaneous plane waves from a B-format signal," in *Proc. IEEE Convention of Electrical & Electronics Engineers in Israel (IEEEI)*, Nov. 2012.

[6] H. Teutsch and W. Kellermann, "Detection and localization of multiple wideband acoustic sources based on wavefield decomposition using spherical apertures," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5276–5279, Mar. 2008.

[7] B. Jo and J. W. Choi, "Direction of arrival estimation using nonsingular spherical ESPRIT," *J. Acoust. Soc. Am.*, vol. 143, pp. EL181–EL187, Mar. 2018.

[8] Q. Huang, L. Zhang, and Y. Fang, "Two-step spherical harmonics ESPRIT-type algorithms and performance analysis," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 26, Sept. 2018.

[9] B. Jo and J. W. Choi, "Nonsingular EB-ESPRIT for the localization of early reflections in a room," *J. Acoust. Soc. Am.*, vol. 144, Sept. 2018.

[10] A. Herzog and E. A. P. Habets, "Eigenbeam-ESPRIT for DOA-vector estimation," *IEEE Signal Process. Lett.*, 2019.

[11] B. Yang, "Projection approximation subspace tracking," *IEEE Trans. Signal Process.*, vol. 43, pp. 95–107, Jan. 1995.

[12] B. Rafaely, *Fundamentals of Spherical Array Processing*, vol. 8. Springer, 2015.

[13] D. P. Jarrett, E. A. P. Habets, and P. A. Naylor, *Theory and Applications of Spherical Microphone Array Processing*. Springer, 2017.

[14] J. Meyer and G. W. Elko, "Spherical microphone arrays for 3d sound recordings," in *Audio Signal Processing for Next-Generation Multimedia Communication Systems* (Y. Huang and J. Benesty, eds.), pp. 67–89, Norwell, MA, USA: Kluwer Academic Publishers, 2004.

[15] A. Björck, *Numerical Methods for Least Squares Problems*. Philadelphia, PA: SIAM, 1996.

[16] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, (South Brisbane, QLD), pp. 5206–5210, Apr. 2015.

[17] D. P. Jarrett, E. A. P. Habets, M. R. P. Thomas, and P. A. Naylor, "Rigid sphere room impulse response simulation: algorithm and applications," *J. Acoust. Soc. Am.*, vol. 132, pp. 1462–1472, Sept. 2012.

[18] *MATLAB 2016b*. Natrick, Massachusetts, US: The MathWorks, Inc.

doi:10.25836/sasp.2019.18

BEAMFORMING WITH DOUBLE-SIDED, ACUSTICALLY HARD PLANAR ARRAYS

Svein Berge
Harpex Ltd
sveinb@harpex.net

ABSTRACT

A disc-shaped baffle with an array of microphones on each surface has recently been proposed as a device for acquiring higher-order ambisonic signals. In this paper, we will study how an array of this type performs with regards to three different beamforming algorithms. The results are quantified through numerical experiments and verified by measurement.

1. INTRODUCTION

Numerous geometries and sensor types have been studied for the purpose of producing microphone arrays with good beamforming properties while at the same time conforming to various practical and economic constraints.

The array studied in this paper is a double-sided array of microphones arranged on a rigid, disc-shaped baffle. There are several motivations for choosing this geometry. Such arrays could be produced at low cost using normal electronics manufacturing techniques. They would be both compact and mechanically robust. They would take up a minimal amount of space in 360 degree video recordings, and could even vanish completely if combined with two half-sphere camera system placed at its center. However, this all is only useful if they also exhibit good acoustical properties.

Arrays of this type have recently been studied for the purpose of acquiring higher-order ambisonic signals [1]. When compared to the more conventional spherical geometry, they have both advantages and disadvantages. The disadvantages stem from the fact that the array has a different symmetry than the desired basis functions, requiring more complex encoding filters that have a lower peak white noise gain, WNG. The advantages stem from the multi-radius nature of these arrays, leading to a wider frequency range. When the resulting noise level is weighted and integrated over the spectrum, the flat array comes out on top for low ambisonic orders and ties with a comparable spherical array at orders 2 and 3.

In this paper, the beamforming performance of the flat array will be studied, following broadly the outline and notation of [2], which studied the same for spherical arrays. That article introduced several beamforming algorithms, some based on ambisonic signals derived from spherical arrays and others using the direct output of the microphone array. Here, we will only study the latter ones, since these have the highest performance.

© Svein Berge. Licensed under a Creative Commons Attribution 4.0 International License (CC BY 4.0). **Attribution:** Svein Berge. "Beamforming with double-sided, acoustically hard planar arrays", 1st EAA Spatial Audio Signal Processing Symposium, Paris, France, 2019.

2. ACOUSTICAL MODEL

We will model the array as a rigid, circular disc-shaped baffle. According to [3], for a plane wave incident at an angle θ_0 with the positive z-axis, such that

$$p_i = \exp\{ik(x\sin\theta_0 + z\cos\theta_0)\}, \quad (1)$$

where k is the wave number, the total field on the top surface of the disc is

$$p_i + p_s = \frac{2}{c}\sum_{m=0}^{\infty}\sum_{n=m}^{\infty}\epsilon_m \frac{i^n}{\widetilde{N}_n^m}\frac{1}{R_{nm}^{3'}(-ika,0)}\times$$
$$S_n^{\mathrm{m}}(-ika,\cos\theta_0)S_n^{\mathrm{m}}(-ika,r/a)\cos m\phi, \quad (2)$$

where p_s is the scattered field, a is the radius of the disc, r is the distance from the center of the disc and $\widetilde{N}$, R and S are defined in [4].

For an incident field equal to one of the eigenfunctions F_n^m of the wave equation in spherical coordinates,

$$p_i = F_n^m = Y_n^m(\theta,\phi)\, j_n(kr), \quad (3)$$

where Y_n^m are the spherical harmonic functions [5] and j_n are the spherical Bessel functions, the scattered field on the top surface simplifies to

$$p_s = C_n^m S_n^m(-ika, r/a)\exp(im\phi), \quad (4)$$

where C_n^m are constants.

The incident field on the bottom surface is equal to that on the top surface, and the scattered field on the bottom surface is opposite that on the top surface.

We express a general incident field as a linear combination of eigenfunctions of the wave equation:

$$p_i = \sum_{n=0}^{\infty}\sum_{m=-n}^{n} a_{nm}(k)Y_n^m(\theta,\phi)\, j_n(kr) \quad (5)$$

where a_{nm} are the coefficients that describe the field. The total field, as sensed by the microphones, is equal to this incident field plus the scattered field

doi:10.25836/sasp.2019.30

$$p_s = \sum_{n=0}^{\infty} \sum_{m=-n}^{n} a_{nm}(k)\, C_n^m \times S_n^m(-ika, r/a) \exp(im\phi). \quad (6)$$

The outer sums in (5) and (6) can be truncated at a finite $N > ka$, since the Bessel functions and the scattered field decrease rapidly in magnitude with n when $n > ka$. The functions are evaluated at the M microphone locations, given by r_i, ϕ_i and $\theta = 0$, bearing in mind that p_s must be negated for microphones on the bottom side of the disc. The result of all this is expressed in this matrix equation, following the notation of [2]:

$$\mathbf{p} = \mathbf{B}\mathbf{a}_{\mathbf{nm}}, \quad (7)$$

where $\boldsymbol{p} = [p(k, r, \theta_1, \phi_1), \cdots, p(k, r, \theta_M, \phi_M)]^T$ is a column vector of length M holding the pressures sampled by the array microphones and the $M \times (N+1)^2$ matrix $\mathbf{B}$ encodes the response of the array to an arbitrary sound field, where each column contains its response to an incident field equal to a single eigenfunction F_n^m. The columns are ordered by n, then by m. The vector $\mathbf{a_{nm}}$ contains the coefficients $a_{nm}(k)$, ordered in the same manner.

3. BEAMFORMERS

Of the beamformers proposed in [2], we will study the "space domain maximum directivity index with optimal alias cancellation" (SMDAC) and the "space domain maximum white noise gain with optimal alias cancellation" (SMGAC) beamformers. We will also study the "sensitivity constrained optimal beamformer" (SCOB) proposed in the context of line arrays in [6]. Apart from a normalization constant, the same beamformers can also be derived from an MVDR formulation [7].

In each case, the beamformer is expressed as a vector of weights $\mathbf{w}$ to be applied to the input signal vector $\mathbf{p}$ in order to produce the output signal y:

$$y(\mathrm{k}) = \mathbf{w}^H \mathbf{p} = \mathbf{w}^H \mathbf{B}\mathbf{a}_{\mathbf{nm}}. \quad (8)$$

Its response A in a given direction (θ_0, ϕ_0) can be found by setting $\mathbf{a_{nm}}$ equal to the coefficients of a planewave, $\mathbf{Y}_0^* = [Y_0^0(\theta_0, \phi_0), \cdots, Y_N^N(\theta_0, \phi_0)]^H$:

$$A(k, \theta_0, \phi_0) = \mathbf{w}^H \mathbf{B}\mathbf{Y}_0^*. \quad (9)$$

3.1 Maximum-directivity beamformer

This beamformer aims to maximize the directivity factor DF of the array, i.e. the output signal power for signals coming from the look direction (θ_l, ϕ_l) relative to the average power for all possible directions of incidence. The expression for the DF is given by

$$DF = 4\pi \frac{\mathbf{w}^H \mathbf{B}\mathbf{Y}_l^* \mathbf{Y}_l^T \mathbf{B}^H \mathbf{w}}{\mathbf{w}^H \mathbf{B}\mathbf{B}^H \mathbf{w}}, \quad (10)$$

The directivity index DI is the directivity factor expressed in dB. The beamformer weights is given by

$$\mathbf{w}^{\mathrm{SMDAC}} = (\mathbf{B}\mathbf{B}^H)^{-1}\mathbf{B}\mathbf{Y}_l^*. \quad (11)$$

It is not normalized (i.e. distortion-free in the MVDR sense), but can be normalized by dividing by its response in the look direction, $A(k, \theta_l, \phi_l)$.

3.2 Maximum white noise gain beamformer

The white noise gain of a beamformer is defined as the improvement in the signal-to-noise ratio in the beamformer output relative to a single sensor in free-field conditions. Its mathematical expression in this context is

$$WNG = \frac{|4\pi A(k, \theta_l, \phi_l)|^2}{\mathbf{w}^H \mathbf{w}}. \quad (12)$$

Its maximum value is achieved with the SMGAC beamformer, defined by

$$\mathbf{w}^{\mathrm{SMGAC}} = \mathbf{B}\mathbf{Y}_l^*. \quad (13)$$

3.3 Sensitivity constrained beamformer

As we will see in the following sections, the SMDAC beamformer cannot be used across the entire spectrum in practice, due to its tendency to amplify sensor noise. However, we may still want a higher directivity than that offered by the SMGAC beamformer, so a compromise between the two might be useful. The sensitivity constrained beamformer provides this through a tradeoff parameter β:

$$\mathbf{w}^{\mathrm{SCOB}} = (\mathbf{B}\mathbf{B}^H + \beta\mathbf{I})^{-1}\mathbf{B}\mathbf{Y}_l^*. \quad (14)$$

For a given WNG, this is the beamformer which optimizes the DI [6]. Or, conversely, for a given DI it optimizes WNG. Setting $\beta = 0$ gives the highest directivity and identical weights to the SMDAC beamformer. Increasing β towards infinity gives the highest WNG. Apart from the constant factor β, which vanishes when the beamformers are normalized, this gives the same weights as the SMGAC beamformer. The optimal value for β will in practice depend on the ratio between sensor noise and ambient noise.

4. NUMERICAL EXPERIMENTS

The array studied here consists of 84 microphones, with 42 placed on either side of the disc as shown in Figure 1.

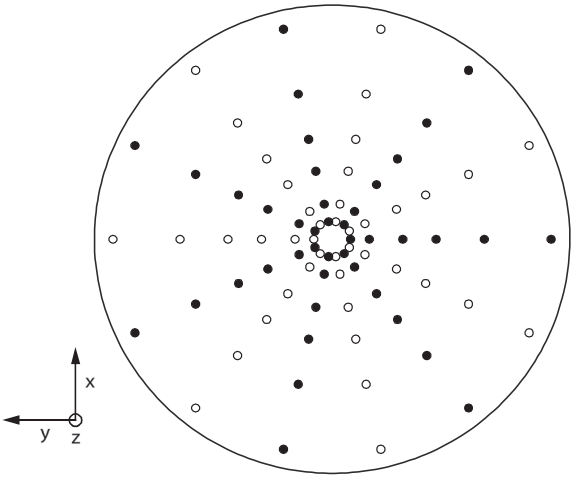


Figure 1. Microphone layout: ○ microphones on top side. ● microphones on bottom side

doi:10.25836/sasp.2019.30

The layout was optimized for producing a 3rd order ambisonic signal with controlled aliasing up to a frequency of 15 kHz [1]. The array has a radius of 85 mm.

The microphones are arranged in six rings of 14 microphones, 7 on each side of the baffle. The radiuses of the rings are given in Table 1.

Ring no.	1	2	3	4	5	6
Radius / mm	6.7	13.1	25.3	37.1	54.2	78.3

Table 1. Microphone ring radiuses.

When used far below the aliasing frequency, the SMDAC beamformer is very sensitive to noise, numerical stability and systematic errors. This makes it unsuitable for use in this frequency range, but it provides a useful upper bound for the directivity index. The SMGAC has no stability problems, but is usually not the optimal choice, since a small reduction in WNG relative to this maximum can usually provide a large increase in directivity index. As a representative of the continuum of beamformers between these two extremes we use the SCOB beamformer with $\beta = 10^{-2}$, which provides a reasonable trade-off between directivity and noise.

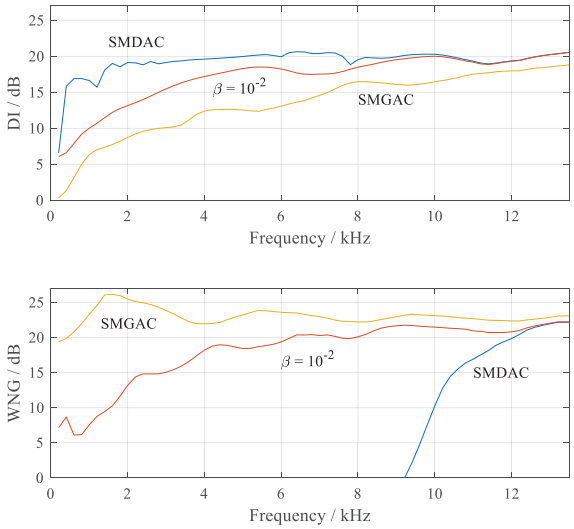


Figure 2. Directivity index and white noise gain for the three beamformers.

Figure 2 shows the directivity index and white noise gain across the spectrum for the three different beamformers when pointed in the $\theta = 0$ direction. At higher frequencies, the differences between the beamformers vanish.

Around 1.5 kHz, the scattering and SMGAC array processing combine favorably to give a maximum WNG of 26 dB, well above the maximum of $10 \log_{10} 84 = 19$ dB for an open array with the same number of microphones. Below about 3 kHz, the SMDAC plot is not reliable due to numerical instability.

Figure 3 shows the beam shapes at one frequency and illustrates how the increase in WNG comes at the cost of a wider main lobe as well as stronger side lobes. As the frequency increases, the number of side lobes will also increase, but their total energy tends to decrease. The slight increase in beam width from SMDAC to, $\beta = 10^{-2}$ provides a dramatic increase in WNG, from -44 dB to 18 dB.

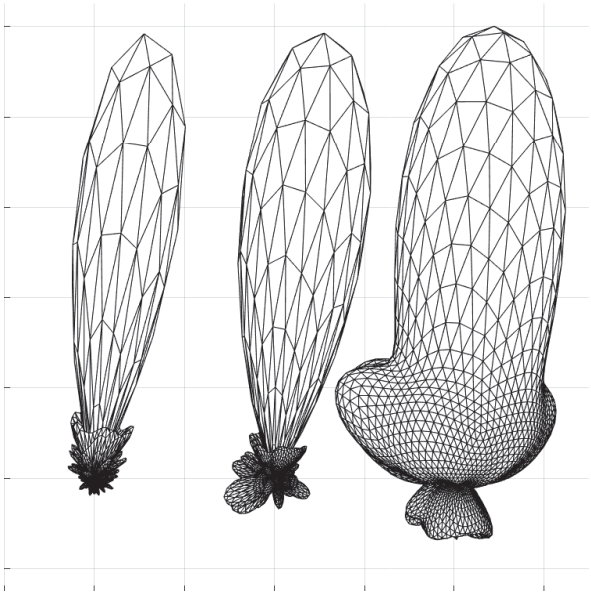


Figure 3. Beam shapes at f = 5 kHz, $\theta_l = 10°$ for the SMDAC, $\beta = 10^{-2}$ SCOB and SMGAC beamformers (left to right). The radial axes in these plots represent linear magnitude.

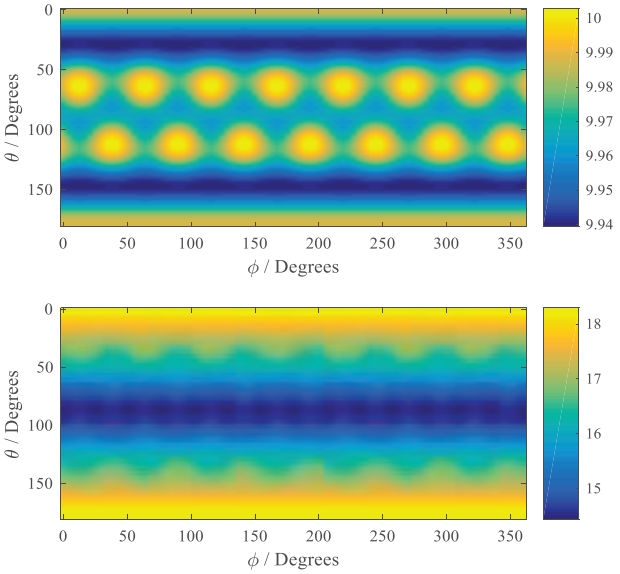


Figure 4. Directivity index for different look directions at f = 1 kHz (top) and f = 5 kHz (bottom). $\beta = 10^{-2}$ SCOB.

As one might expect from an array which is not spherically symmetrical, its directivity pattern is also not symmetrical. At low frequencies, the directivity is isotropic for all practical purposes. Above about 1 kHz, the directivity is highest along the z axis, as seen in Figure 4.

doi:10.25836/sasp.2019.30

5. EXPERIMENTAL VERIFICATION

The correctness and practical applicability of the theoretical results is verified using a physical device consisting of 84 IM69D130 microphones placed on a 1.6 mm thick printed circuit board made from the fiberglass-based laminate FR-4 (Figure 5). The circuit board is further laminated between a 1.0 mm sheet of pressboard, a 0.3 mm polystyrene foil and 0.5 mm polyester fabric on either side. The total thickness is 5.5 mm, and the outer radius is 85 mm. The microphones are placed according to the model in the previous section. The microphones are connected to a computer via a USB interface.

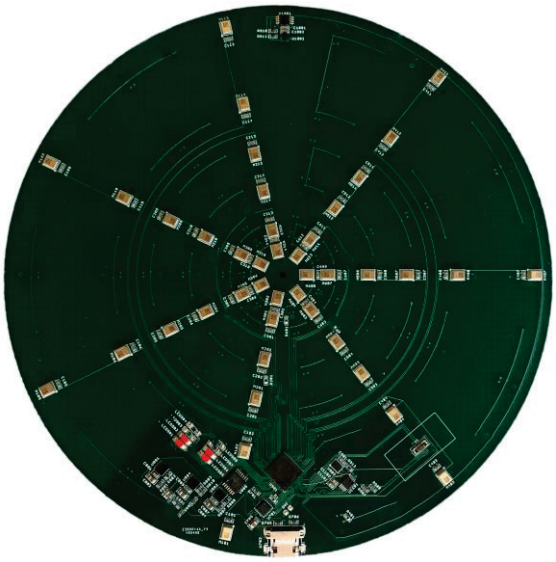


Figure 5. Experimental measurement device

The device is suspended from the ceiling using a 3 mm brass tube with a length of 1.5 m (Figure 6). Power and signals are sent through wires inside this tube. One end of the tube is connected to the edge of the circuit board and the other end is attached to an angle gauge, allowing measurements to be taken at a series of rotations about the device's x axis. The tube is stabilized with guy wires to prevent lateral movement of the device during rotation.

A loudspeaker is placed 2 m away from the device. The loudspeaker consists of two concentric drivers which were driven separately and combined in post-processing with a crossover frequency of around 8 kHz. The loudspeaker enclosure is axisymmetric and airtight. The room is not anechoic. Apart from the loudspeaker, the device and their supports, there are no objects or structures within a volume with less than 1 m additional path length. The impulse response measurements should therefore be free from external reflections up to 2.9 ms, and only the first 1.5 ms are used in the following. The impulse responses are measured according to the methods in [8] for every 5° of θ from -90° to 90°.

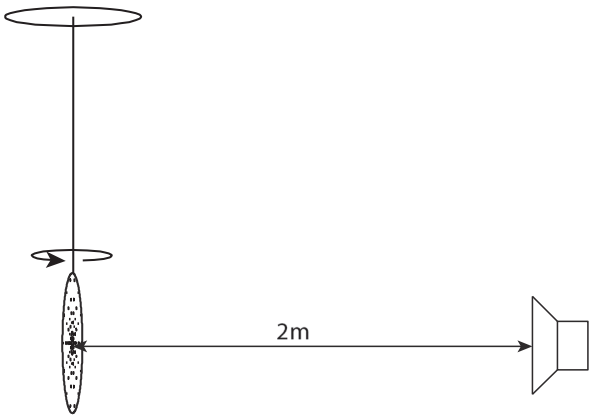


Figure 6. Measurement setup (figure not to scale).

The noise spectrum of the microphones was measured by recording the output of one microphone with a blocked acoustic port. The absolute level of the spectrum was shifted to match the A-weighted noise level of 25 dB (A) given in the device's data sheet [9]. Using this, it was possible to calculate the A-weighted equivalent noise levels of the three beamformers shown in Table 2.

Beamformer	SMDAC	$\beta = 10^{-2}$	SMGAC
Equiv. self noise	113 dB (A)	13 dB (A)	3 dB (A)

Table 2. Noise level, $\theta = 0°$.

Since the measurement setup only allows rotation about the array's x axis, we can only directly measure the beam patterns in the y-z plane, as in Figure 8. To access the beam pattern in the horizontal plane, we measure the response of the array in one horizontal direction and calculate the response as the look direction of the beamformer is rotated around the horizon, resulting in Figure 7.

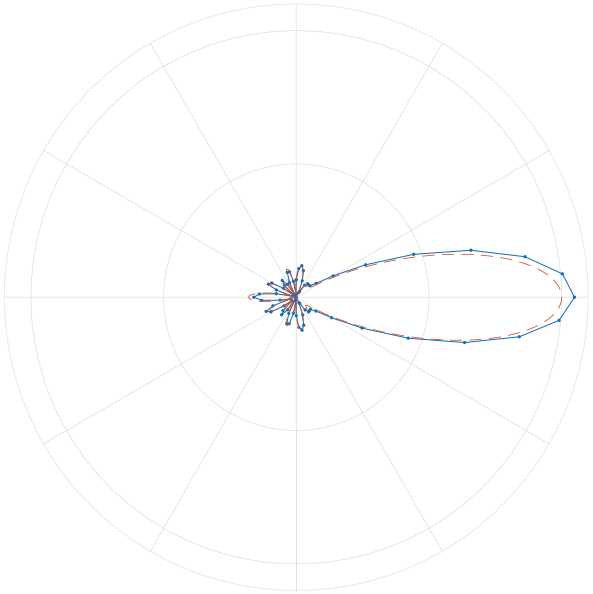


Figure 7. Modeled (– –) and measured (—) response, SCOB $\beta = 10^{-2}$ in the horizontal plane at f = 5 kHz. Radius is linear magnitude response.

doi:10.25836/sasp.2019.30

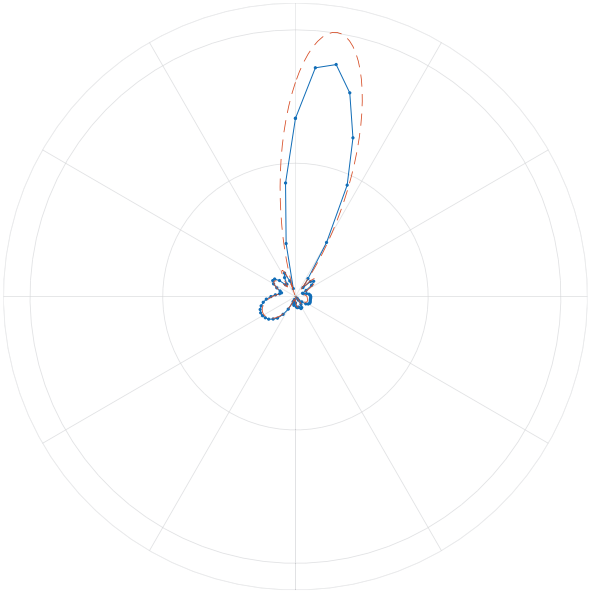


Figure 8. Modeled (– –) and measured (—) response in the y-z plane at f = 5 kHz, $\beta = 10^{-2}$ and $\theta_l = 10°$. Radius is linear magnitude response.

The same analysis that produced Figure 8 is repeated across the frequency range to produce an overview of the frequency dependency of the polar patterns. The result is shown in Figures 9 and 10. In this last figure, the front response is normalized. The measured data for the lowest frequencies (< 1 kHz) may not be reliable due to the truncated impulse response measurement method.

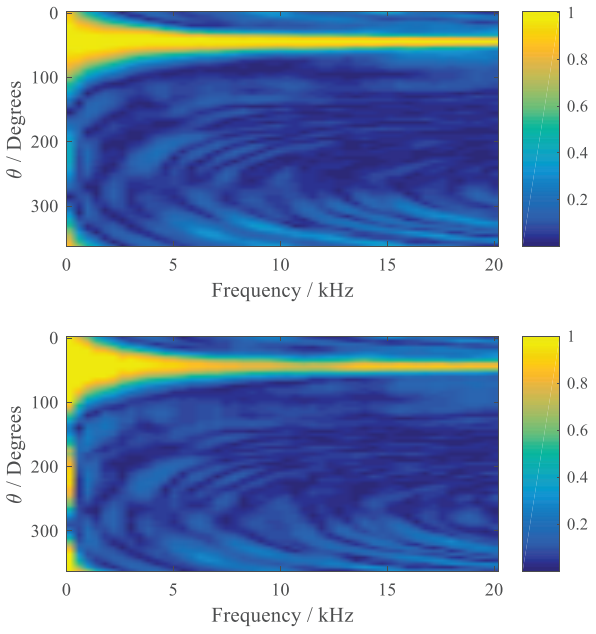


Figure 9. Modeled (top) and measured (bottom) response in a vertical plane at different frequencies, $\beta = 10^{-2}$ and $\theta_l = 45°$.

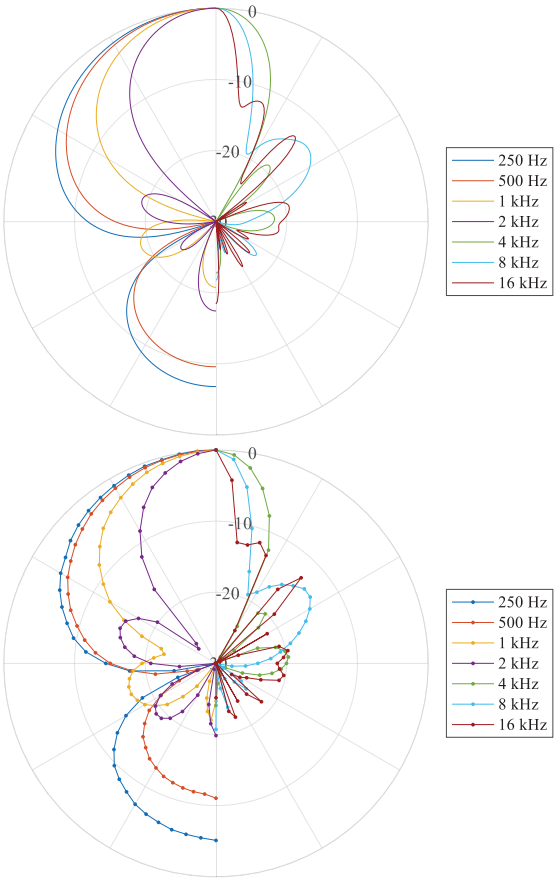


Figure 10. Modeled (top) and measured (bottom) response in the y-z plane, $\beta = 10^{-2}$ and $\theta_l = 0°$. Radius is dB magnitude response.

6. CONCLUSION

Some of the beamforming techniques that were developed for spherical arrays and line arrays also work well with double-sided disc arrays. The array was originally developed with 3rd order ambisonics in mind. However, if the ultimate goal is to perform beamforming, the direct approach studied here gives significantly better results. Array-agnostic 3rd order ambisonic beamformers are limited to 12 dB directivity index and, for arrays of this type, about 0 dB white noise gain, while the SCOB beamformer can provide a WNG of 15-20 dB with the same directivity index. It was already known that the SMDAC beamformer works best above the array's aliasing limit. Because of the multi-radius nature of the arrays studied here, the transition between no aliasing and full aliasing takes place over a much wider range than for a spherical array. The SMDAC beamformer can only be used in the upper part of this range and above.

Since double-sided disc arrays have a non-isotropic scattering function, they work better in some directions than in others. The beamformers described here take optimal advantage of any scattering that takes place. Particularly at medium to high frequencies, this effect provides a higher directivity and / or white noise gain along the z axis than in other directions. This means that for applica-

doi:10.25836/sasp.2019.30

tions where the approximate direction of arrival can be predicted before setting up the microphone, this may be a better option than a spherical array, whereas in applications where an isotropic response is required, a spherical array should be used.

7. REFERENCES

[1] Svein Berge, "Acoustically Hard 2D Arrays for 3D HOA," in *Audio Engineering Society Conference: 2019 AES International Conference on Immersive and Interactive Audio*, 2019.

[2] David Lou Alon and Boaz Rafaely, "Beamforming with Optimal Aliasing Cancellation in Spherical Microphone Arrays," *IEEE/ACM trans. on audio, speech, and language processing*, vol. 24, no. 1, January 2016.

[3] Frederick B. Sleator, "The Disc," in *Electromagetic and acoustic scattering by simple shapes*.: Michigan Univ Ann Arbor Radiation Lab, 1970.

[4] Carson Flammer, *Spheroidal wave functions*.: Stanford University Press, 1957.

[5] George B. Arfken and Hans J. Weber, *Mathematical Methods For Physicists*, 5th ed.: Academic, 2001.

[6] Henry Cox, Robert M. Zesking, and Theo Kooij, "Practical Supergain," *IEEE Trans. on Acoustics, speech, and signal processing*, vol. ASSP-34, no. 3, 1986.

[7] Shefeng Yan, Haohai Sun, U. Peter Svensson, Xiaochuan Ma, and Jens M. Hovem, "Optimal modal beamforming for spherical microphone arrays," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 2, 2011.

[8] Angelo Farina, "Simultaneous measurement of impulse response and distortion with a swept-sine technique," in *Audio Engineering Society Convention 108*, 2000.

[9] Infineon, "IM69D130 - High performance digital XENSIV MEMS microphone," Datasheet 2017.

 doi:10.25836/sasp.2019.30

VIRTUAL ACOUSTIC RENDERING BY STATE WAVE SYNTHESIS

Esteban Maestre
CAML, McGill University
esteban@music.mcgill.ca

Gary P. Scavone
CAML, McGill University
gary@music.mcgill.ca

Julius O. Smith
CCRMA, Stanford University
jos@ccrma.stanford.edu

ABSTRACT

We present State Wave Synthesis (SWS), a framework for the efficient rendering of sound traveling waves as exchanged between multiple directional sound sources and multiple directional sound receivers in time-varying conditions. We introduce a mutable state-space system modal formulation through which input/output matrices change size and coefficients in time-varying conditions, and perceptually motivated designs are possible. SWS enables the accurate simulation of frequency-dependent source directivity and receiver directivity, provides means for simulating frequency-dependent attenuation of propagating waves, and allows an alternative realization through which state variables are treated as propagating waves.

1. INTRODUCTION

Among the challenges of Virtual Reality object-based spatial audio, directionality of sound sources and listeners remains a capstone. In fact, the quest for efficient techniques to model and simulate Head-Related Transfer Functions (HRTF) has arguably been among the most popular in the field. In virtual environments that allow for multiple moving sources, a *classic* form for interactive HRTF simulation requires a database of directional impulse or frequency responses, run-time interpolation of responses, and a time- or frequency-domain convolution engine. Given the predominantly minimum-phase nature of HRTF [1], improved interpolation quality is often achieved if interaural excess-phase is modeled separately as a function of direction. With this form, which is employed in two recently reported systems [2, 3], it is straightforward to swap between personalized HRTF databases [4, 5]; however, one run-time interpolation-convolution channel is required per source wavefront. To reduce the computational needs for each channel, techniques have been studied to simulate HRTF via IIR filters [6], leading to one independent IIR filter per wavefront. Despite the reduction, run-time interpolation of generic IIR filter coefficients is a hard task and leads to artifacts. A number of linear decomposition methods have been used to separate HRTF into parallel basis functions, involving PCA or ICA [7], SVD [8], or Spherical Harmonics (SH) [9, 10]. In turn, the favorable qualities of these parallel models have led to interactive rendering schemes where the parallel basis functions are fixed in number and each of them is implemented by convolution. Following the common-pole observations advanced in [11], state-space models of HRTF have been proposed where size and coefficients of the input matrix are fixed during simulation [12, 13]; though reporting lower computational costs than convolution systems, the size of the input matrix increases with spatial resolution as each system input is statically associated to a fixed direction; also, the methods proposed for joint estimation of transition and input matrices do not allow perceptually-motivated designs as those employed for IIR binaural filters [6]. With respect to efficient simulation of source directivity, some recent works have instead focused on proposing systematic approaches to obtain model-based sound radiation fields using the Equivalent Source Method (ESM) and SH [14]. By low-order SH applied to ESM and pre-computations on fixed virtual scenes [15], frequency responses incorporating source and listener directivity up to around 1 kHz can be generated at rates of 10-15 Hz.

In the context of virtual acoustic simulation relying on traveling wave rendering as dictated by path-tracing methods we present State Wave Synthesis (SWS), a time-domain, convolution-free framework for rendering traveling waves between moving sources and listeners. We introduce a mutable state-space modal formulation that enables simulating directivity of both sources and listeners in terms of input and output matrices of time-varying size and coefficients. We work by first identifying eigenvalues and then constructing time-varying input/output matrix models, thus allowing perceptually-motivated frequency resolutions. The most basic form of SWS comprises three main components: mutable state-space models in modal form used to represent sound source and receiver objects, wave propagators in the form of fractional delay lines and attenuation elements, and input and output mapping functions that facilitate interfacing objects to wave propagators. While the most relevant strength of SWS is its ability to efficiently simulate the directivity of sources and receivers, it also offers means to simulate frequency-dependent attenuation of propagating waves by manipulating the state variables of source object models, and enables a convenient reformulation through which sound propagation is simulated by propagating state variables of source object models.

2. SOURCE AND RECEIVER OBJECTS

Each sound source or receiver object is modeled as a time-varying dynamical system and simulated in terms of a *mutable* state-space model. We use the term *mutable state-space model* to refer to a state-space model for which both its number of input or output variables and their associated input or output vectors mutate dynamically. In such a con-

doi:10.25836/sasp.2019.05

text, we impose that models correspond to strictly proper transfer functions expressible in state-space modal form. Then we write the discrete-time update relation of a mutable state-space model of an object as

$$\begin{aligned} \underline{s}[n+1] &= \underline{\lambda} \odot \underline{s}[n] + \sum_{p=1}^{P} \underline{b}^p[n] x^p[n] \\ y^q[n] &= \underline{c}^q[n]^T \underline{s}[n], \end{aligned} \tag{1}$$

where n is the time index, "$\odot$" denotes element-wise vector multiplication, $\underline{s}[n]$ is a vector of M state variables, $\underline{\lambda}$ is the vector of M system eigenvalues, $x^p[n]$ is the p-th input (a scalar) of those existing at time n, $\underline{b}^p[n]$ is its corresponding length-M vector of input projection coefficients, $y^q[n]$ is a q-th system output (a scalar) obtained as a linear projection of the state variables, and $\underline{c}^q[n]$ is the corresponding length-M vector of output projection coefficients. We refer to the q-th product

$$\underline{c}^q[n]^T \underline{s}[n] \tag{2}$$

as the q-th *output mapping*. The output mapping coefficient vectors enable the projection of the state space of the model onto the output space. We refer to the p-th product

$$\underline{b}^p[n] x^p[n] \tag{3}$$

as the p-th *input mapping*. The input mapping coefficient vectors enable the projection of the input space onto the state space of the model. Note that, as opposed to the classic, fixed-size matrix-based state-space model notation, here we resort to a more convenient vector notation because both the number of inputs or outputs and the coefficients in their corresponding projection vectors (i.e., the numerator coefficients of an equivalent parallel system) are allowed to change (*mutate*) dynamically. In a first basic form, source objects are represented as mutable state-space models for which their outputs are mutable but their inputs are non-mutable (i.e., a fixed number of inputs and input projection coefficients); conversely, receiver objects are represented as mutable state-space models for which their inputs are mutable but their outputs are non-mutable (i.e., a fixed number of outputs and output projection coefficients). However, the framework allows object models for which both inputs and outputs are mutable. Models are constructed by first identifying or defining a set of eigenvalues, and then designing time-varying input/output matrix models via *input/output mapping functions* as outlined below. This two-step procedure allows perceptually-motivated designs.

3. INPUT AND OUTPUT MAPPING FUNCTIONS

Input and output mapping functions enable the mutability of inputs or outputs of a source or receiver object in terms of a number of dynamically changing coordinates associated to those inputs or outputs. For example, the coordinates associated to an input of a sound receiver object may refer to the position or orientation from which the receiver object is excited by a sound wave. As the key elements in mapping functions, *projection models* enable the approximation of the distribution of input and output projection coefficient values over the space of input and output coordinates of an object. Mapping functions employ such projection models to estimate projection coefficients. Without loss of generality, for now we associate input mapping functions to receiver object models and output mapping functions to source object models.

The input mapping function S^+ of a receiver object estimates the input vector $\underline{b}^p[n]$ of projection coefficients corresponding to its p-th input, as a function of an input projection model $\mathcal{B}$ and a vector $\underline{\beta}^p[n]$ of input coordinates. This can be expressed as

$$\underline{b}^p[n] = S^+\big(\mathcal{B}, \underline{\beta}^p[n]\big). \tag{4}$$

Analogously, the output mapping function S^- of a source object estimates the vector $\underline{c}^q[n]$ of output projection coefficients corresponding to the q-th system output, as a function of an output projection model $\mathcal{C}$ and a vector $\underline{\zeta}^q[n]$ of output coordinates. This is expressed as

$$\underline{c}^q[n] = S^-\big(\mathcal{C}, \underline{\zeta}^q[n]\big). \tag{5}$$

Mapping functions may be devised from arbitrary designs or from discrete measurement data. Data-driven construction of projection models enables transforming discrete sets of known projection coefficients (designed, or estimated from measurements) into multivariate continuous functions over the space of the input or output coordinates of an object. This allows having a continuous, smooth time-update of projection coefficients while, for instance, objects change positions or orientations. Notwithstanding the possibility of formulating projection models by way of elaborate modeling methods (e.g., regression in terms of basis functions of different kinds), interpolation of known coefficient vectors may remain cost-effective because only look-up tables are needed. If constrained by memory it should be possible to encode the distribution of projection coefficients via SH instead of storing look-up tables, but this would incur an additional computational cost during regression.

4. WAVE PROPAGATION

A sound wave propagating from the q-th output $y^q[n]$ of a source object to the p-th input $x^p[n]$ of a receiver object may suffer frequency-independent delay induced by the traveled distance, distance-related frequency-independent attenuation induced by wave propagation along the path, and frequency-dependent attenuation due to obstacle interactions (e.g., reflection, transmission, diffraction) or other attenuation causes along the path. A simple model representing these three phenomena can be formulated as

$$x^p[n] = \alpha[n]\, y^q\big[n - l[n]\big] * \chi[n], \tag{6}$$

where $\alpha[n]$ is the scaling factor associated to frequency-independent attenuation, $l[n]$ is the number of delay samples corresponding to the traveled distance given the wave

doi:10.25836/sasp.2019.05

propagation speed and the simulation sampling rate, and $\chi[n]$ is the impulse response of a linear system that models the accumulated frequency-dependent attenuation characteristic $\chi(\omega)$ derived from any obstacle interactions happening along the path or other attenuation causes. Note that $\alpha[n]$, $l[n]$, and $\chi[n]$ may all be time-varying. If the effect of the system characterized by $\chi[n]$ is approximated by a low-order digital filter, a wave propagator responds to a more *classic* structure, for which time-varying frequency-attenuation should be handled by retrieving and/or slewing the coefficients of such low-order filter. Below we will see that, thanks to the state-space formulation, frequency-dependent attenuation can be approximated by scaling the state variables of directional source object models.

5. FREQUENCY-DEPENDENT ATTENUATION VIA STATE ATTENUATION

As an alternative to designing, implementing, and managing digital filters with the sole purpose to model frequency-dependent attenuation due to propagation or obstacle interactions along a path, if the eigenvalues of an object model are conveniently distributed and their associated low-pass (positive real eigenvalue), band-pass (complex-conjugate eigenvalue pair), or high-pass (negative real eigenvalue) components cover representative frequency bands, it is possible to approximate the propagation-induced frequency-dependent attenuation by simply attenuating its state variables at the time of projection. We describe this by using a source object as an example.

For a sound wave traveling from the q-th output of a source object model to the p-th input of a receiver object model, a desired propagation-induced frequency-dependent attenuation characteristic $\chi(\omega)[n]$ may be approximated in terms of a length-M vector $\underline{\gamma}^q[n] = (\gamma_1^q[n], \ldots, \gamma_m^q[n], \ldots, \gamma_M^q[n])$ of source state variable attenuation factors, applied prior to the q-th output projection. In this way, the q-th sound wave $y^q[n]$ departing from the source object model already incorporates the desired attenuation for the path. The new output update relation becomes

$$y^q[n] = \underline{c}^q[n]^T \big(\underline{\gamma}^q[n] \odot \underline{s}[n]\big) \tag{7}$$

where coefficients in $\underline{\gamma}^q[n]$ are real and satisfy $\gamma_m^q[n] \leq 1 \, \forall m = 1, \ldots, M$. With this, the wave propagator relation reduces to

$$x^p[n] = \alpha[n]\, y^q\big[n - l[n]\big]. \tag{8}$$

The system component in charge of estimating the vector $\underline{\gamma}^q[n]$ of state attenuation coefficients is what we refer to as a *state attenuation function*. To estimate the state attenuation coefficients, the state attenuation function may simply sample the desired frequency-dependent attenuation characteristic $\chi(\omega)$ at M frequencies ω_m each corresponding to the natural frequency associated to the m-th eigenvalue of the model. Besides enabling a simplification of the overall implementation provided that a directional source object model presents a convenient set of eigenvalues, this allows the path attenuation function to be easily updated at run-time while the path is still active, e.g., to enable time-varying attenuation properties.

6. STATE WAVE FORM

Through an alternative formulation that we name the *state wave form*, the exact same results can be obtained if eliminating the delay lines of the sound wave propagators and instead treating the source object state variables as propagating waves. To outline this important, yet simple transformation of the system, let us first assume that frequency-dependent attenuation is approximated via low-order digital filters at the end of each propagator delay line. By attending to Equation (1), it should be noted that a traveling wave departing from an object only depends on the state variables of the object model and the vector of coefficients involved in the output projection. Once the output projection is executed, the resulting wave is fed into a fractional delay line for propagation. Let us assume that a sound-emitting object model is feeding a traveling wave $y^q[n]$ into a fractional delay line, and let us refer to the output signal of such delay line as the *delayed traveling wave* signal $d^q[n] = y^q\big[n - l[n]\big]$ with $l[n]$ the fractional sample delay of the line. The delayed traveling wave signal $d^q[n]$ can be expressed in terms of the state variable vector $\underline{s}[n]$ and the output projection coefficient vector $\underline{c}^q[n]$ via

$$d^q[n] = \underline{c}^q\big[n - l[n]\big]^T \underline{s}\big[n - l[n]\big], \tag{9}$$

where $l[n]$ is the delay line length, and $\underline{c}^q\big[n - l[n]\big]$ and $\underline{s}\big[n - l[n]\big]$ are delayed versions of the output coefficient vector and the state variable vector respectively. Since the delayed coefficient vector $\mathbf{c}^q\big[n - l[n]\big]$ can be estimated by the output mapping function given the delayed output coordinates $\underline{\zeta}^q\big[n - l[n]\big]$, i.e.,

$$\underline{c}^q\big[n - l[n]\big] = S^-\Big(\mathcal{C}, \underline{\zeta}^q\big[n - l[n]\big]\Big), \tag{10}$$

the delayed traveling wave $d^q[n]$ can be obtained via

$$d^q[n] = S^-\Big(\mathcal{C}, \underline{\zeta}^q\big[n - l[n]\big]\Big)\underline{s}\big[n - l[n]\big] \tag{11}$$

in terms of *delayed state variables* $\underline{s}\big[n - l[n]\big]$ and *delayed coordinates* $\underline{\zeta}^q\big[n - l[n]\big]$. Thus, instead of propagating traveling waves as emitted by source objects, the system propagates the state variables and the position/orientation coordinates of those objects. If we now assume that frequency-dependent attenuation is approximated via state attenuation, each traveling wave arriving to a sound-receiving object is simply obtained by tapping from the emitting object state variable delay lines and coordinate delay lines at a desired position $l[n]$, and sequentially performing the operations for state attenuation and output projection. This form, which may incur an increase in the cost induced by fractional delay, can be advantageous in diverse application contexts because it eliminates the need for allocating and deallocating delay lines associated to individual propagation paths as traced in dynamically changing scenes. Anecdotally, with this formulation

it could be possible to approximate wave dispersion by tapping at slightly different positions during fractional delay interpolation of the delayed state variables.

7. EXAMPLE MODELS

To provide some simple and illustrative example models, we choose a three-dimensional spatial domain where sound waves propagate as spherical waves radiated from a sound emitting object, propagating in any outward direction from a sphere of finite size representing the object. The direction and position of wave emission by the source are encoded by two angles of three-dimensional spherical coordinates, and constant radius. An equivalent assumption is made for the receiver object: sound waves from a source at a given distance are received from any direction, encoded by two spherical coordinate angles. To portray the presumably more difficult case of modeling real objects as opposed to numerical models of objects, we choose a real acoustic violin as the source object, and a real human body as the receiver object. We demonstrate frequency-dependent attenuation via state attenuation of the acoustic violin model.

7.1 Source object: acoustic violin

An acoustic violin was measured in a low-reflectivity chamber, exciting the bridge with an impact hammer and measuring the sound pressure with a microphone array. The transversal horizontal force exerted on the bass-side edge of the bridge was measured, and defined as the only input of the system. As for the outputs, the resulting sound pressure signals were measured at 4320 positions on a centered spherical sector surrounding the instrument, with radius 0.75 meters from a chosen center coinciding with the middle point between the bridge feet. The spherical sector being modeled covered approximately 95% of the sphere. Each measurement position corresponds to a pair (θ, φ) of angles in the vertical polar convention, conforming the output coordinates on a two-dimensional rectangular grid of $60 \times 72 = 4320$ points. Such a grid represents the uniform sampling of a two-dimensional euclidean space whose dimensions are θ and φ. To design the mutable state-space model of the violin, we first impose minimum-phase and then estimate 58 eigenvalues over a warped frequency axis while jointly accounting for all responses. We then define the input matrix of a corresponding *classic* state-space model as a sole, length-58 vector of ones, and estimate the 4320×58 output matrix by solving a least-squares minimization problem. The solution to this problem equivalently provides estimations for $M = 58$ matrices of size 60×72, with each m-th matrix representing the spherical distribution of output projection coefficient values corresponding to the m-th eigenvalue. From the estimated output matrix, we construct an output mapping function that performs bilinear interpolation of output coefficients over the two-dimensional space of output coordinates, i.e., over the two-dimensional space of angles (θ, φ). To demonstrate the behavior of the source model at run-time, we slew the output coordinates of an outgoing wave as captured by an ideal microphone lying on the sphere surrounding the source object. Assuming ideal excitation of the violin bridge, we simulate a continuous linear motion of the ideal microphone on the sphere, from initial position at ($\theta = 0.69$ rad, $\varphi = 4.71$ rad) to a final position at ($\theta = -1.48$ rad, $\varphi = -0.52$ rad). To illustrate the quality and smoothness of the achieved result, in Figure 1 we compare the measured responses (nearest-neighbor) and the modeled responses as obtained from bilinear interpolation of output projection coefficients.

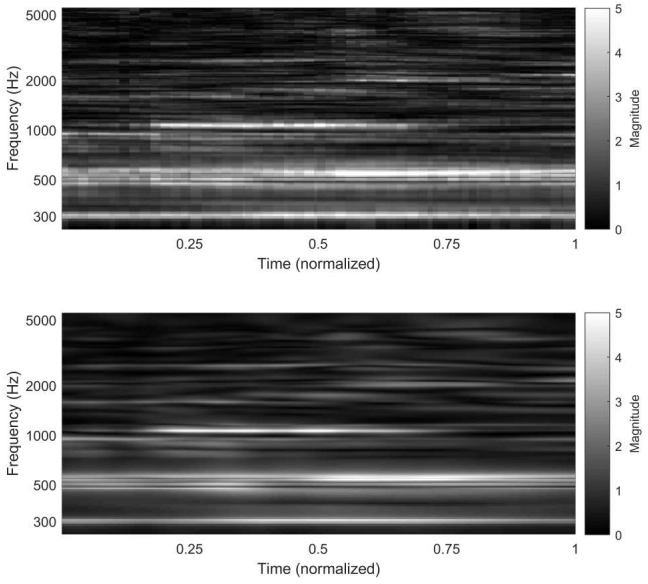


Figure 1. Example modeling (M = 58) of a real acoustic violin as a source object with a spherical sector of 0.75-meter radius as output space expressed in vertical polar coordinates, and the bridge force signal as its only input. Input-output magnitude frequency response as obtained from exciting the model in time-varying conditions: continuous linear motion of an ideal microphone on the sphere, from initial position at ($\theta = -0.69$ rad, $\varphi = -0.34$ rad) to final position at ($\theta = 5.06$ rad, $\varphi = 1.39$ rad). Top graph: nearest-neighbor measurement; bottom graph: model.

7.2 Receiver object: HRTF

To demonstrate the modeling of a receiver object we choose a human body sitting in a chair, as represented by a high-spatial resolution head-related transfer function set of the CPIC dataset [16]. The data used here comprises 1250 responses obtained from measuring the left in-ear microphone signal during excitation by a loudspeaker located at 1250 unevenly distributed positions on a head-centered spherical sector of 1-meter radius. The spherical sector being modeled covers approximately 80% of the sphere. Each of the 1250 excitation positions corresponds to a pair (θ, φ) of azimuth and elevation angles in a two-dimensional space of input coordinates, expressed in the inter-aural polar convention. Following a warped-frequency design procedure analogous to that employed for the violin, we first design a classic state-space model of 1250 inputs, 36 state variables, and one output. From such a model, we first smooth and uniformly upsample the coordinate input space to form $M = 36$ matrices each of $64 \times 64 = 4096$ coefficient values and again choose bilin-

doi:10.25836/sasp.2019.05

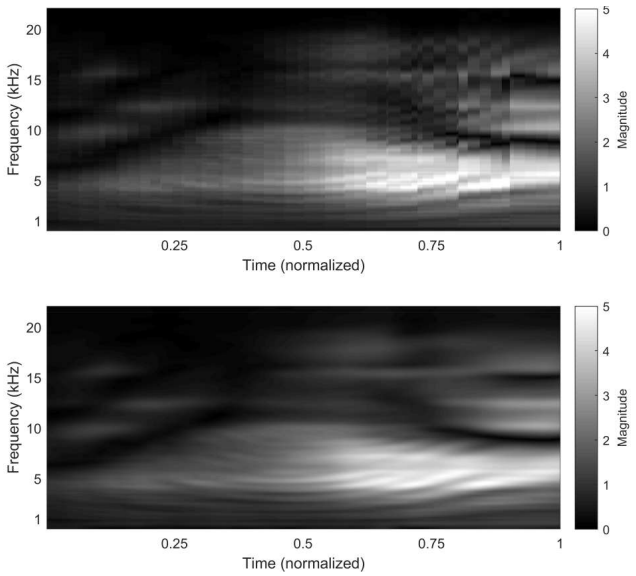


Figure 2. Example modeling (M = 36) of a sitting human body as a receiver object with a head-centered spherical sector of 1-meter radius as input space expressed in inter-aural polar coordinates, and the left in-ear microphone signal as its only output. Input-output magnitude frequency response as obtained from exciting the model in time-varying conditions: continuous linear motion of an ideal source on the sphere, from initial position at ($\theta = 0.69$ rad, $\varphi = 4.71$ rad) to a final position at ($\theta = -1.48$ rad, $\varphi = -0.52$ rad). Top graph: nearest-neighbor measurement; bottom graph: model.

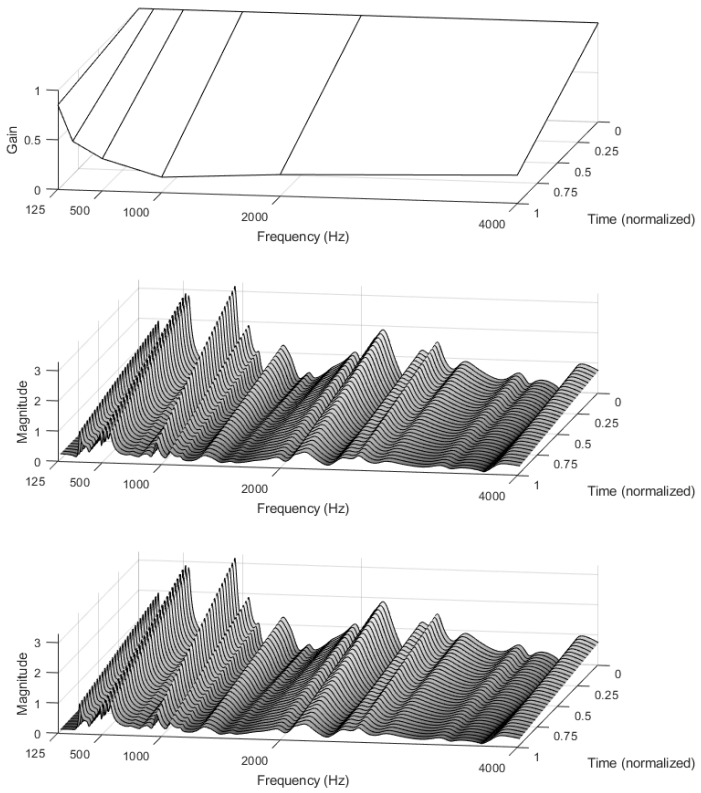


Figure 3. Example modeling of frequency-dependent attenuation via state variable attenuation of a violin model with $M = 58$. For 32 consecutive steps, we modify the frequency-dependent attenuation of an emitted sound wave towards direction ($\theta = -0.34$ rad, $\varphi = 1.39$ rad) by linearly fading from no attenuation to that caused by reflection off a cotton carpet as provided in material tabulated data. Top graph: gain as derived by the attenuation characteristic; middle graph: attenuation simulated via frequency-domain convolution; bottom graph: attenuation simulated via state variable attenuation.

ear interpolation as the input mapping function.

We synthesize the input-output frequency response as obtained from exciting the model in time-varying conditions. For 512 consecutive steps, we modify the input coordinates of an incoming wave as emitted by an ideal source lying on the sphere surrounding the receiver object. We simulate a continuous linear motion of the ideal source on the sphere, from initial position at ($\theta = 0.69$ rad, $\varphi = 4.71$ rad) to a final position at ($\theta = -1.48$ rad, $\varphi = -0.52$ rad). To illustrate the quality and smoothness of the achieved result, in Figure 2 we again compare the measured responses (nearest-neighbor) and the model responses (bilinear interpolation of input coefficients). In the context of binaural rendering, two collocated receiver object models similar to the one demonstrated here could be used, one for each ear. Since the collocated models share position and orientation, the direction coordinates of the incoming traveling wave can be used for both input projection functions. In such context, the required inter-aural time difference can be simulated by tapping from two different positions of the delay line of the incoming wave, and feeding each obtained signal to its respective collocated receiver model.

7.3 Frequency-dependent attenuation

To demonstrate frequency-dependent attenuation via state variable attenuation, we employ the acoustic violin object model described above. Given eight gains $\chi(\omega_k)$, with $\chi(\omega_k) \leq 1 \, \forall k$ and $\omega_1 \cdots \omega_k \cdots \omega_8$ corresponding to octave band frequencies, to estimate the gain required for each of the $M = 58$ state variables of the acoustic violin model (each with natural frequency ω_m), the state attenuation function performs linear interpolation of known gains $\chi(\omega_k)$. We illustrate the time-varying frequency-dependent attenuation by linearly fading, through 32 consecutive steps, between no attenuation (i.e., $\chi_m = 1 \, \forall m = 1, \cdots, M$) and the attenuation caused by a reflection off cotton carpet. This is illustrated in Figure 3, where we compare the results obtained via state attenuation to those obtained by magnitude-only frequency-domain convolution of the departing wave with a response constructed with our scheme.

8. OUTLOOK

We introduced SWS, a time-domain, convolution-free framework that uses a mutable state-space modal formulation for the efficient simulation of both source and receiver directivity in interactive virtual acoustic rendering applications. With our state-space structure, input/output matrices change size and coefficients in time-varying conditions while allowing efficient diagonal forms where, given a state-space order, the computational cost is independent of the spatial resolution. Moreover, our two-stage de-

sign process allows for perceptually-motivated designs via warped-frequency eigenvalue identification as a first step. The framework, which scales well with the number of wavefronts and also enables the simulation of frequency-dependent attenuation, offers a flexible quality-cost trade-off in terms of the order of the state-space models, and allows for a realization where state variables of source objects may be treated as propagating waves.

Straightforward extensions are possible. Though we presented SWS as an early-field renderer, discrete directional components of a simulated diffuse field could be rendered as independent directional wavefronts. The proposed mutable state-space modal representation readily allows for scattering behavior via collocated input and output coordinate spaces and mutability of inputs and outputs, which could be combined with source or receiver behaviors into state-shared hybrid object models. Under linear conditions and relying on the diagonalized modal form of state-space models, it should be possible to dedicate the (mutable) inputs of source objects to serve as a coupling mechanism between a virtual acoustic rendering system and animation-driven rigid-body simulations or physical models. Other attractive routes include simulating effects like near-field and diffraction, or even non-linear source or receiver behaviors in terms of mutable eigenvalues.

Though the provided examples were picked to demonstrate the effectiveness of SWS in accurately simulating highly-directive objects while ensuring smoothness under interactive operation, developments are ongoing and a thorough study of computational cost versus perceptual quality is still required for a fair comparison to other systems. Nevertheless, it remains apparent that simplicity (state-space modal form), efficiency (input and output projections by look-up tables and short vector-scalar multiply-adds), and flexibility (perceptually-motivated frequency designs, state-space order selection, state wave form) make SWS an attractive and resourceful framework for virtual acoustic rendering in diverse application contexts, with some potential for parallel hardware implementations.

9. REFERENCES

[1] J. Nam, M. Kollar, and J. S. Abel, "On the minimum-phase nature of head-related transfer functions," in *AES 125th Convention*, 2008.

[2] D. Poirier-Quinot and B. F. G. Katz, "The anaglyph binaural audio engine," in *AES 144th Convention*, 2018.

[3] M. Cuevas-Rodriguez, L. Picinali, D. Gonzalez-Toledo, C. Garre, E. de la Rubia-Cuestas, L. Molina-Tanco, and A. Reyes-Lecuona, "3d tune-in toolkit: An open-source library for real-time binaural spatialisation," *PLoS ONE*, vol. March, 2019.

[4] C. Hoene, I. C. P. Mejia, and A. Cacerovschi, "Mysofa - design your personal hrtf," in *AES 142nd Convention*, 2017.

[5] A. Meshram, R. Mehra, H. Yang, E. Dunn, J.-M. Franm, and D. Manocha, "P-hrtf: Efficient personalized hrtf computation for high-fidelity spatial sound," in *IEEE Int. Symposium on Mixed and Augmented Reality*, 2014.

[6] J. Huopaniemi and J. O. Smith, "Spectral and time-domain preprocessing and the choice of modeling error criteria for binaural digital filters," in *16th AES Int. Conf. on Spatial Sound Reproduction*, 1999.

[7] V. Larcher, J.-M. Jot, J. Guyard, and O. Warusfel, "Study and comparison of efficient methods for 3d audio spatialization based on linear decomposition of hrtf data," in *AES 108th Convention*, 2000.

[8] J. S. Abel and S. H. Foster, "Method and apparatus for efficient presentation of high-quality three-dimensional audio including ambient effects," *US Patent and Trademark Office*, US5802180, 1998.

[9] B. Rafaely and A. Avni, "Interaural cross correlation in a sound field represented by spherical harmonics," *Journal of the Acoustical Society of America*, vol. 127:2, pp. 823–828, 2009.

[10] M. Noisternig, T. Musil, A. Sontacchi, and R. Holdrich, "3d binaural sound reproduction using a virtual ambisonic approach," in *IEEE Int. Symposium on Virtual Environments, Human-Computer Interfaces and Measurement Systems*, 2003.

[11] Y. Haneda, S. Makino, Y. Kaneda, and N. Kitawaki, "Common-acoustical-pole and zero modeling of head-related transfer functions," *IEEE Trans. on Speech and Audio Processing*, vol. 7:2, pp. 188–196, 1999.

[12] P. Georgiou and C. Kyriakakis, "A multiple input single output model for rendering virtual sound sources in real time," in *IEEE Conf. on Multimedia and Expo*, 2000.

[13] N. H. Adams and G. H. Wakefield, "State-space synthesis of virtual auditory space," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 16(5), 2008.

[14] D. L. James, J. Barbic, and D. K. Pai, "Precomputed acoustic transfer: output-sensitive, accurate sound generation for geometrically complex vibration sources," *ACM Trans. on Graphics*, vol. 35:1, pp. 987–995, 2006.

[15] R. Mehra, L. Antani, S. Kim, and D. Manocha, "Source and listener directivity for interactive wave-based sound propagation," *IEEE Trans. on Visualization and Computer Graphics*, vol. 20:4, pp. 495–503, 2014.

[16] V. R. Algazi, R. O. Duda, D. M. Thompson, and C. Avendano, "The cipic hrtf database," in *IEEE Workshop on Applications of Signal Processing to Audio and Electroacoustics*, 2001.

doi:10.25836/sasp.2019.05

SIMULATIVE INVESTIGATION OF REQUIRED SPATIAL SOURCE RESOLUTION IN DIRECTIONAL ROOM IMPULSE RESPONSE MEASUREMENTS

Johannes Klein, Michael Vorländer
Institute of Technical Acoustics, RWTH Aachen University, Aachen, Germany

ABSTRACT

In classical and standardized room acoustic measurements, the directivity of sources is either not considered, or specified to be omni-directional for inter-measurement comparability. These room impulse response measurements allow for the analysis of certain aspects of an acoustic scene. Parameters such as Early Decay Time attempt to reflect the listening impression, while others such as the reverberation time serve as physical descriptors and for the evaluation of rooms for different purposes.

The explicit non-consideration of the source and receiver directivity waives an abundance of information, which is essential for a complete description of the spatial composition of an acoustic scene, e.g. for an auralization. On the one hand, a more complete description raises the degree of realism in auralizations, presumably intensifying the immersion of human listeners. On the other hand, the freedom of arbitrary source-receiver directivity combinations opens up the possibility of directional acoustical analysis like scanning single reflection paths within a room. For the measurement of directional room impulse responses several measurement methods and instruments have been implemented in the past. Most of them aim either at fast measurements or at measurements of a high spatial resolution. The compromise of obtaining a sufficiently high resolution in an acceptable time usually is often disregarded due to the missing definition of a sufficiently high resolution.

Before the development of suitable methods and instruments, the importance of directivity in an acoustical scene has to be determined. This entails the question if and up to which complexity the measurement of directional room impulse responses offers an advantage for room representations in auralizations and for parametric room acoustic descriptions. A sufficiently high resolution can be defined in many ways, either in regard to just noticeable differences for human listeners, or objectively regarding the effects on technical parameters. This work investigates the effect of the source directivity resolution on room acoustical parameters. Since real measurements contain too many uncontrollable influences such as time variances and sources that can radiate the required directivity with a sufficient precision do not yet exist, this investigation is done using room acoustic simulations.

First, the modeling of a suitable artificial directivity will be explained. The spatial resolution will be denoted by the corresponding spherical harmonic order. The goal is a directivity with a minimum beam width for each given spherical harmonic order without strong side lobes. This characteristic represents the worst case for each resolution. The generated source directivity will then be used in hybrid ray tracing and image source room acoustic simulations in two rooms of different size and acoustic property. This approach allows a more generally valid statement about the impact of the source directivity. The results will be discussed using the impact on objective room acoustic parameters as an indicator for the required spatial source resolution. Subjective parameters will be considered as an outlook toward the impact on the human perception. The findings are meant to aid the design of measurement instruments for directional room impulse response measurement in reasonable measurement times with a sufficiently high spatial resolution.

© Johannes Klein, Michael Vorländer. Licensed under a Creative Commons Attribution 4.0 International License (CC BY 4.0). **Attribution:** Johannes Klein, Michael Vorländer. "Simulative Investigation of Required Spatial Source Resolution in Directional Room Impulse Response Measurements", 1st EAA Spatial Audio Signal Processing Symposium, Paris, France, 2019.

1. INTRODUCTION

The measurement of *Room Impulse Responses* (RIRs) with arbitrary source directivity promises an advantage for room representations in auralizations and for parametric room acoustic descriptions. However, before the development of suitable measurement methods and instruments for such *Directional Room Impulse Response* (DRIR) measurements, it has to be investigated if and up to which complexity [1] they are required.

This study compares the results of simulations in several acoustic scenes with the complexity of the source directivity as the variable. The definition of complexity is ambiguous, hence the worst case for each level of complexity is defined and used in the simulations. The resulting DRIRs are analyzed regarding their temporal structure and resulting room acoustic parameters to evaluate a threshold for the required source directivity complexity either in each room or over a wide range of rooms.

[1] *complexity* refers to the spatial detail and resolution of a directivity.

doi:10.25836/sasp.2019.24

2. FUNDAMENTALS

Some definitions and fundamentals are required to facilitate the understanding of the investigation.

2.1 Spherical Harmonic Base

The complex *Spherical Harmonic* (SH) base functions can be used to decompose any azimuth (φ) and elevation (ϑ) angle-dependent spatial function $f(\vartheta, \varphi)$ on the unit sphere into its fundamentals. This study uses the complex SH base functions $Y_n^m(\vartheta, \varphi)$ containing the associated Legendre functions P_n^m [1] defined as [2] [2]

$$Y_n^m(\vartheta, \varphi) = \sqrt{\frac{(2n+1)}{4\pi}\frac{(n-m)!}{(n+m)!}} \cdot P_n^m(\cos(\vartheta)) \cdot \mathrm{e}^{\mathrm{j}m\varphi}. \tag{1}$$

2.2 Spherical Wave Spectrum

The transformation of a spatial function into its fundamentals results in a coefficient function $\hat{f}_{nm}$, denoting the share of each fundamental in *order* (n) and *degree* (m). This coefficient function is called *Spherical Wave Spectrum* (SWS). The *Spherical Harmonic Transform* (SHT) is defined as [2]

$$\hat{f}_{nm} = \mathcal{S}\{f(\vartheta, \varphi)\} = \oint_{S^2} f(\vartheta, \varphi) \cdot \overline{Y_n^m(\vartheta, \varphi)} \mathrm{d}\Omega. \tag{2}$$

The operation yields a precise coefficient function, if the integral can be solved in an exact way. The *Inverse Spherical Harmonic Transform* (ISHT) is defined as [2]

$$f(\vartheta, \varphi) = \mathcal{S}^{-1}\{\hat{f}_{nm}\} = \sum_{n=0}^{\infty}\sum_{m=-n}^{n} \hat{f}_{nm} \cdot Y_n^m(\vartheta, \varphi). \tag{3}$$

A finite summation limit can be applied (*truncation*), resulting in a less precise spatial function.

2.3 Dirac on a Sphere

The completeness relation for the SH is using the Kronecker delta δ as [2]

$$\sum_{n=0}^{\infty}\sum_{m=-n}^{n} Y_n^m(\vartheta, \varphi)\overline{Y_n^m(\vartheta', \varphi')} = \delta^{(\vartheta', \varphi')}(\vartheta, \varphi). \tag{4}$$

Analogous to the Dirac impulse in the time domain, the function differs from zero only in the one case of $\vartheta = \vartheta', \varphi = \varphi'$. Comparing Eq. (4) with Eq. (3) yields the SWS coefficients for the spatial Dirac on the unit sphere

$$\delta_{nm}^{(\vartheta', \varphi')} = \overline{Y_n^m(\vartheta', \varphi')}. \tag{5}$$

The spatial Dirac integrates to unity [2], and the sifting property holds [2]. The spatial Dirac is the "1"-element of the spatial convolution, which for SWS is defined as [4]

$$\mathcal{S}\{f(\Omega) * g(\Omega)\} = 2\pi\sqrt{\frac{4\pi}{2n+1}}\hat{f}_{nm}\hat{g}_{n0}. \tag{6}$$

with the coefficients $\hat{g}_{n0}$ only defined for a degree $m = 0$.

[2] This definition omits the explicit mention of the Condon-Shortley phase $(-1)^m$, due to its inclusion in the associated Legendre function [3].

2.4 Exterior Problems

In exterior problems all sources are confined to a volume with a radius r_0. The exterior is source-free, and the Sommerfeld radiation condition is met. With the Hankel function of the second kind h_n, the extrapolation of SWS coefficients $\hat{c}_{nm}$ from r_0 to a radius r is defined as [2]

$$\hat{c}_{nm}(kr) = \hat{c}_{nm}(kr_0)\frac{h_n(kr)}{h_n(kr_0)}. \tag{7}$$

2.5 Order-Far-Field

The extrapolation frequency-dependently annihilates high orders. While passing through the near-field, this effect is stronger for coefficients for high orders at low frequencies compared to coefficients of low orders, distorting the directivity. Starting at an *order-far-field* distance, these order-relative differences in attenuation become negligible.

2.6 Order Truncation

Order truncation in the SWS domain introduces artifacts. Their magnitude is determined by the highest retained order and the truncation method. Predictors of substantial orders for a specific combination of wave number k and source radius r can be found in [5–7], the most common being the *kr-limit* [8]

$$n_{\mathrm{trunc}} = \lfloor kr \rfloor. \tag{8}$$

The coefficients can be cut-off (*rect* function) or faded out using for example a Hann function [9].

2.7 Vibrating Polar Cap

The vibrating polar cap describes a uniform surface vibration of a north-pole sphere section limited by the elevation aperture angle α. The SWS coefficients for the vibration are [6]

$$\hat{u}_{nm} = \begin{cases} \sqrt{\pi}(1-\beta), & n = 0, \\ \delta_{m0}\sqrt{\frac{\pi}{2n+1}}\left[P_{n-1}(\beta) - P_{n+1}(\beta)\right], & n > 0, \end{cases} \tag{9}$$

with $\beta = \cos(\alpha)$ and the Legendre functions P_{n-1} and P_{n+1}. The acoustic radiation impedance with the speed of sound c and the density ρ_0 connects the vibration coefficients and the sound pressure at the boundary surface [6]

$$\hat{p}_{nm}(r_0) = -\mathrm{j}\rho_0 c\frac{h_n(kr_0)}{h_n'(kr_0)}\hat{u}_{nm}(r_0). \tag{10}$$

The coefficients $\hat{p}_{nm}$ can be rotated to any (ϑ', φ') using the spherical convolution and the Dirac given in Eqs. (5) and (6). The coefficients can be extrapolated to larger radii using Eq. (7). The Hankel functions introduce a frequency-dependent order distribution, further diminishing coefficient values for high orders in low frequencies.

doi:10.25836/sasp.2019.24

3. ARTIFICIAL SOURCE DIRECTIVITY GENERATION

The source directivity needs to be scalable and represent the worst case for each level of complexity. Here, the worst case is defined as directivity with a frequency dependent minimum-width main-lobe. The directivity is artificially generated in the SWS domain to ensure a consistent worst case directivity. Two methods for generating such a directivity are discussed in this section.

3.1 Acoustic Beam

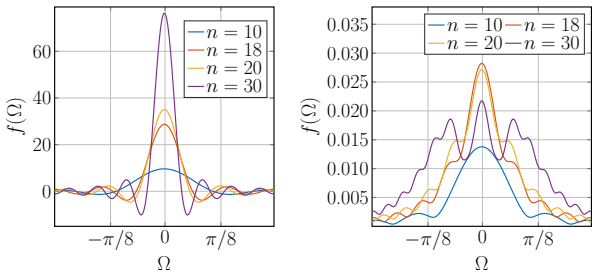
The acoustic beam is based on the spatial Dirac. It can be freely positioned using the spherical convolution in Eq. (6). It can simply be extrapolated using Eq. (7) (cf. Fig. 1).

3.2 Spherical Cap

The spherical cap is widely used as a model for transducers on a spherical array [2, 6, 10, 11]. In its original form it describes the same actively vibrating surface at all frequencies, contradicting the actual behavior of loudspeakers [12, 13] and the worst case defined above.

3.3 Order Limitation

The orders allowed by the kr-limit still contain a sub-range of high orders that is deformed significantly stronger on the way to the order-far-field than the lower orders. The amplitude of the Dirac functions on the source surface in Fig. 1a rises and the beam-width narrows monotonously with the truncation order, this is not true after the extrapolation in Fig. 1b. The peak value of the Dirac with a truncation order of 20 is lower than that of order 18, while side-lobes start to rise. At a truncation order of 30 the side-lobe amplitude almost reaches that of the further attenuated main-lobe.



(a) Source surface. **(b)** Extrapolation.

Figure 1: Amplitude of n-truncated Dirac functions over angular distance to the center-axis direction. Extrapolation $0.15\text{m} \rightarrow 100\text{m}$ for $f = 22.05\text{kHz}$.

Using the first zero-crossings of the first derivative of the Dirac peak value over the truncation order, a stricter *kr-root-limit* limit can experimentally be determined. This prevents the unwanted deformation of the order-far-field

$$n_{\sqrt{kr}_{1,s,e}} = \begin{cases} \lfloor s \cdot \sqrt{k \cdot r_s} - 1 \rfloor, & k \cdot r_s \geq \sqrt{\frac{1}{s}} \\ e, & k \cdot r_s < \sqrt{\frac{1}{s}}. \end{cases} \quad (11)$$

4. SIMULATION SETTINGS AND PARAMETERS

This section presents the methodology, the specific source directivity, and the room models for the simulations.

4.1 Simulation Methods

A hybrid model of ray tracing and image sources, implemented in the room acoustic simulation software RAVEN [14, 15] is used for all simulations. The image source order is set to 2 and $1 \cdot 10^6$ ray tracing particles are deployed. The ray tracing reflection pattern and the impulse response generation Poisson sequence are fixed. The directivity is defined as magnitude spectra at 65160 points of a regular horizontal and vertical 1° sampling scheme, allowing for the representation of an order of 50. The monaural receiver directivity is chosen to be omnidirectional. The binaural simulations use the head-related transfer-functions of the ITA artificial head [16] and at the same receiver center positions.

The surface material information and the geometrical models for the simulation are taken from the Benchmark for Room Acoustical Simulation (BRAS) database [17]. The material coefficients are fitted[3] for all models to match the measured mean T30 values in the BRAS database at the ANSI center frequencies [18] using the ITA-Toolbox[4] [19] and RAVEN. Below 50Hz the absorption values are spline-interpolated.

4.2 Simulated Sources

An acoustic beam source with a source radius $r_\text{s} = 0.15\text{m}$, limited to $\sqrt{kr}_{1,\frac{64}{25},1}$, smoothed with a Hann function and extrapolated to $r = 100\text{m}$ is used for the simulations. Only the absolute value is transferred to the simulation. Simulations are done for a main-lobe alignment along all 6 spatial axes. Fig. 2 shows the directivity for selected frequencies.

4.3 Simulation Rooms

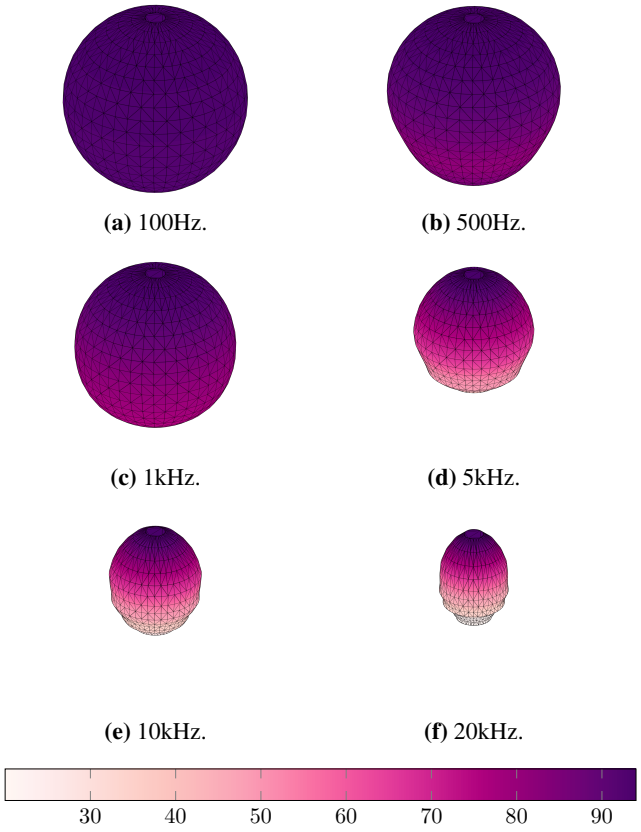
Two different rooms from the BRAS database are used as simulation models. The simulation and measurement data is obtained for all permutations of 2 loudspeaker and 5 receiver positions. The measured room characteristics from the BRAS database are briefly presented in this section.

4.3.1 Medium Concert Hall (Scene 10)

The medium concert hall has a surface area of 2720m^2 and a volume of 3319m^3. It is designed for classical music concerts. The mean T30 is 1.184s, resulting in a Schroeder frequency of 37.77Hz. The reverberation time is very constant for the frequency range up to 1kHz and decays to higher frequencies. The T30 reverberation times for all positions show a mean standard deviation of $s = 0.05\text{s}$ between 100Hz and 10kHz. The mean broadband EDT over all positions is 0.8545s. The energy ratios indicate a high clarity and depend strongly on the measurement position.

[3] The starting values for the fitting process are the *fitted estimates* included in [17].

[4] Git commit SHA dffe6d84524aca91cec8bcbe55f582fa8bb421e8.



(a) 100Hz. **(b)** 500Hz. **(c)** 1kHz. **(d)** 5kHz. **(e)** 10kHz. **(f)** 20kHz.

Figure 2: Balloon plot of the absolute values of acoustic beam source directivity aligned with +z.

4.3.2 Large Auditorium (Scene 11)

The large auditorium has a $5811\mathrm{m}^2$ surface area and a volume of $8658\mathrm{m}^2$. It is designed for lectures. The mean T30 over all positions is 1.956s, resulting in a Schroeder frequency of 30.06Hz. The T30 decays over the frequency with a slight elevation around 200Hz. The T30 reverberation times for all positions show a mean standard deviation of $s = 0.064$s between 100Hz and 10kHz. The mean broadband EDT over all positions is 1.3367s. The energy ratios show a high variance.

5. RESULTS

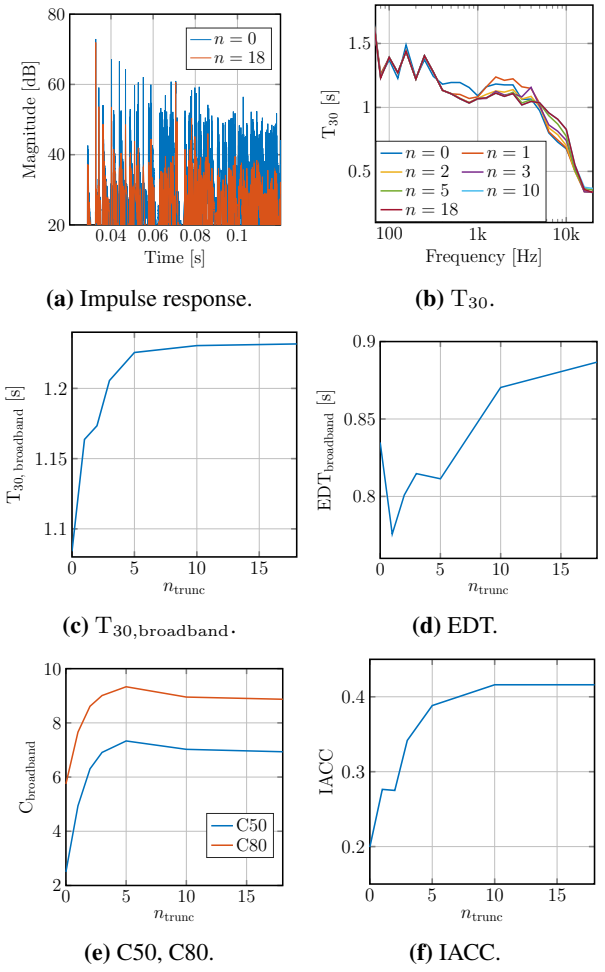
For reasons of brevity, only a small selection of the simulation results is presented. The measurement point numbering and axes correspond to the BRAS database.

5.1 Medium Concert Hall (Scene 10)

The results for simulations with a forward (+x) facing source at position 1 and a receiver at position 3 are shown in Fig. 3. Fig. 4 shows the results for a simulation with an upward (+z) facing source at the same position and a receiver at measurement position 2.

The RIR in Fig. 3a shows, that the measurement position 3 is still well in the beam of the forward facing source, while measurement position 2 is clearly not for the upward facing beam, as seen in Fig. 4a. Especially the behavior of the energy ratios in Figs. 3e and 4e is affected by this difference. While for measurement position 3 the clarity rises with the directivity due to the disappearance of room reflections, it falls with the order for position 2 as soon as it is not within the beam aperture.

The behavior of the other parameters is similar for both simulations, while the values differ strongly. The broadband T30 is prolonged by 0.13s or 12% for the first simulation and by 0.5 or 50% for the second (cf. Figs. 3c and 4c). The EDT rises about 0.1s or 12% for simulation 1 and by 0.5s or 50% for simulation 2 (cf. Figs. 3d and 4d). The *Inter-Aural Cross Correlation* (IACC) rises by roughly 0.23 for both simulations. The narrow-band T30s in Figs. 3b and 4b show some variance for frequencies up to 10kHz. Most curves reach a saturation around the truncation order 10, only the EDT changes up to the truncation order 18.



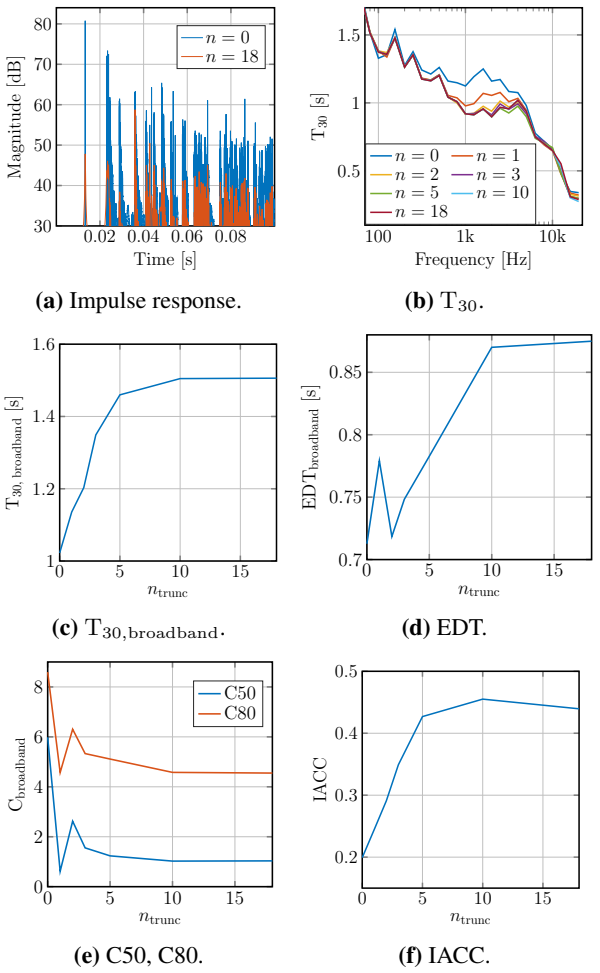
(a) Impulse response. **(b)** T_{30}. **(c)** $\mathrm{T}_{30,\mathrm{broadband}}$. **(d)** EDT. **(e)** C50, C80. **(f)** IACC.

Figure 3: Simulation results. Scene 10, source position 1, direction +x, measurement point 3.

5.2 Large Auditorium (Scene 11)

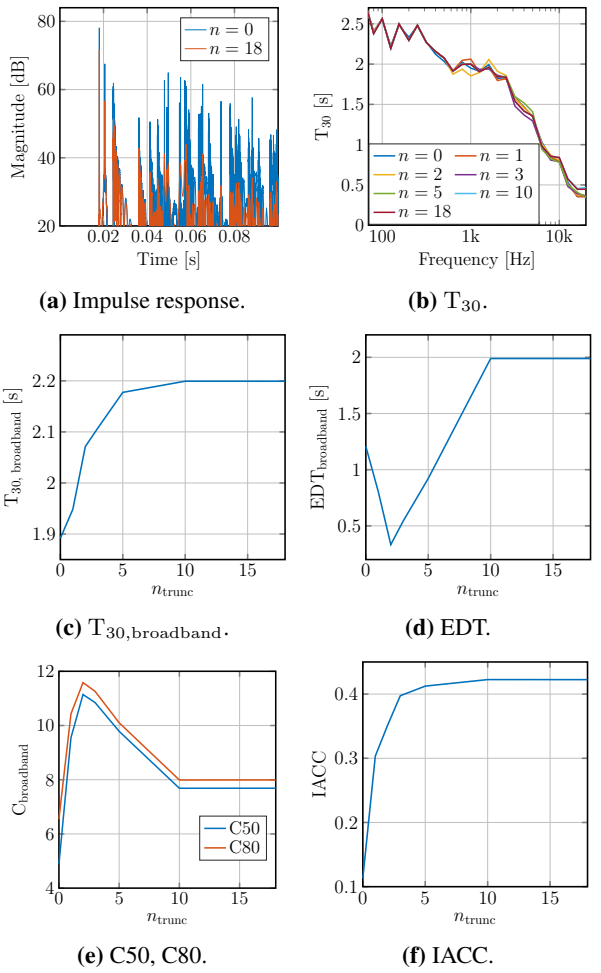
Fig. 5 shows the simulation results for the large auditorium with an acoustic beam source in forward (+x) orientation at source position 1, recorded with a receiver at position 4. Fig. 6 shows a backward (-x) oriented beam source at the same position, recorded by a receiver at position 3.

doi:10.25836/sasp.2019.24



(a) Impulse response. **(b)** T_{30}. **(c)** $T_{30,broadband}$. **(d)** EDT. **(e)** C50, C80. **(f)** IACC.

Figure 4: Simulation results. Scene 10, source position 1, direction +z, measurement point 2.



(a) Impulse response. **(b)** T_{30}. **(c)** $T_{30,broadband}$. **(d)** EDT. **(e)** C50, C80. **(f)** IACC.

Figure 5: Simulation results. Scene 11, source position 1, direction +x, measurement point 4.

While position 4 is still well in the forward beam (cf. Fig. 5a), the direct sound is nearly missing at position 3 for the backward beam, as seen in Fig. 6a. This is also without any doubt an audible effect.

Both T30 simulations in Figs. 5b and 6b are again only slightly affected. For the simulation with the forward facing source, the T30 rises by 0.3s (16%, Fig. 5c), the EDT by 1.5s (300%, Fig. 5d), and the IACC by 0.31. For the simulations with the backward facing source, the T30 rises by 0.3s (15%, Fig. 6c), the EDT by 0.6s (38%, Fig. 6d), and the IACC by 0.4.

For the measurement point in the source beam, the energy ratios rise as long the point is within the aperture (cf. Fig. 5e), while any other truncation order than 0 for the backward facing source means a lesser clarity, as seen in Fig. 6f.

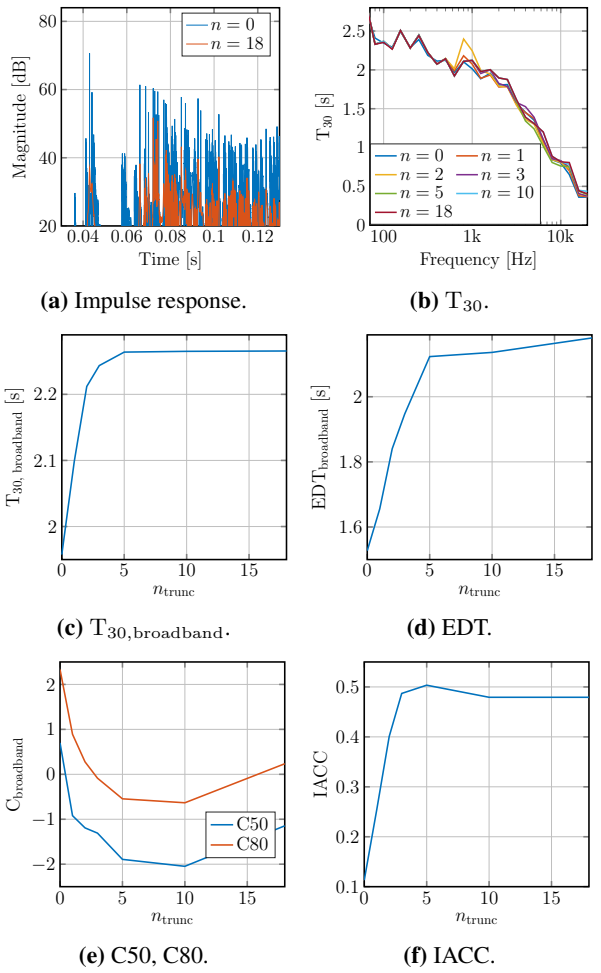
6. CONCLUSION

The simulations use the worst-case directivity for a common source with a radius $r_s = 0.15$m. The source radius dictates the respective truncation order at every specific frequency.

The simulations show a significant directivity impact on the RIR and some room acoustic parameters. The temporal structure of the RIR is affected by the absence of reflections and the direct sound level by the orientation of the source, which can be considered as easily audible. Especially the C50 and C80 energy ratios yield information about the relative positioning of the source beam and the receiver position. The IACC generally profits from a more diffuse field without many prominent reflections. The T reverberation times are not as affected by the directivity, since they do not regard the direct sound and the early reflections. The EDT rises by 50%-300% due to a changing source directivity. Since the EDT is considered a measure for the perception of reverberation, this is a clear indicator that the source directivity matters for RIR measurements and the perception of acoustic scenes.

In the simulations in can be seen that most parameters show an order-saturation, since the beam width changes minimally and so do the parameters. The effect is illustrated for the RIR of the simulation in the large auditorium, for the backward facing source and receiver position 3 in Fig. 7. The RIR correlation finds a saturation above a truncation order of 10. This should also be taken into account during the design of instruments for DRIR measurements.

doi:10.25836/sasp.2019.24



(a) Impulse response. **(b)** T_{30}.

(c) $T_{30,broadband}$. **(d)** EDT.

(e) C50, C80. **(f)** IACC.

Figure 6: Simulation results. Scene 11, source position 1, direction -x, measurement point 3.

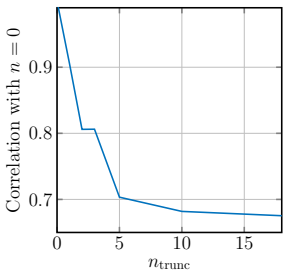


Figure 7: Scene 11. Correlation of the simulated RIRs for several orders with RIR for $n = 0$.

7. REFERENCES

[1] M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions*. Courier Corporation, 1964.

[2] E. G. Williams, *Fourier Acoustics*. Academic press, 1999.

[3] G. B. Arfken, H.-J. Weber, and F. E. Harris, *Mathematical Methods for Physicists*. Elsevier, 2013.

[4] J. R. Driscoll and D. M. Healy, "Computing Fourier Transforms and Convolutions on the 2-Sphere," *Advances in Applied Mathematics*, vol. 15, no. 2, 1994.

[5] N. A. Gumerov and R. Duraiswami, "Computation of Scattering from N Spheres Using Multipole Reexpansion," *J. Acoust. Soc. Am.*, vol. 112, no. 6, 2002.

[6] B. Rafaely, *Fundamentals of Spherical Array Processing*. Springer Berlin Heidelberg, 2015.

[7] M. Müller-Trapet, M. Pollow, and M. Vorländer, "Spherical Harmonics as a Basis for Quantifying Scattering and Diffusing Objects," in *Proc. Forum Acusticum*, (Aalborg, Denmark), 2011.

[8] R. Duraiswami, D. N. Zotkin, and N. A. Gumerov, "Interpolation and Range Extrapolation of HRTFs," in *Proc. IEEE ICASSP*, (Montreal, Canada), 2004.

[9] A. V. Oppenheim and R. W. Schafer, *Discrete-time signal processing*. Pearson, 2010.

[10] M. Pollow and G. K. Behler, "Variable Directivity for Platonic Sound Sources Based on Spherical Harmonics Optimization," *Acta Acustica united with Acustica*, vol. 95, no. 6, 2009.

[11] F. Zotter and R. Höldrich, "Modeling Radiation Synthesis with Spherical Loudspeaker Arrays," in *Proc. ICA*, (Madrid, Spain), 2007.

[12] F. J. M. Frankort, "Vibration Patterns and Radiation Behavior of Loudspeaker Cones," *J. Audio Eng. Soc*, vol. 26, no. 9, 1978.

[13] W. Klippel and J. Schlechter, "Distributed Mechanical Parameters of Loudspeakers, Part 1: Measurements," *J. Audio Eng. Soc*, vol. 57, no. 7/8, 2009.

[14] D. Schröder and M. Vorländer, "RAVEN: A Real-Time Framework for the Auralization of Interactive Virtual Environments," in *Proc. Forum Acusticum*, (Aalborg, Denmark), 2011.

[15] D. Schröder, *Physically Based Real-Time Auralization of Interactive Virtual Environments*. Logos Berlin, 2011.

[16] A. Schmitz, "Ein neues digitales Kunstkopfmeßsystem," *Acustica*, vol. 81, no. 4, 1995.

[17] L. Aspöck, F. Brinkmann, D. Ackermann, S. Weinzierl, and M. Vorländer, "BRAS - Benchmark for Room Acoustical Simulation," 2019. `http://dx.doi.org/10.14279/depositonce-6726.2`.

[18] American National Standards Institute, "ANSI S1.11-2004: Specification for Octave-Band and Fractional-Octave-Band Analog an Digital Filters," 2004.

[19] M. Berzborn, R. Bomhardt, J. Klein, J.-G. Richter, and M. Vorländer, "The ITA-Toolbox," in *Proc. DAGA*, (Kiel, Germany), 2017.

doi:10.25836/sasp.2019.24

ASSESSING THE ANISOTROPIC FEATURES OF SPATIAL IMPULSE RESPONSES

Benoit Alary
Acoustics Lab
Dept. of Signal Processing and Acoustics
Aalto University
Espoo, Finland
Benoit.Alary@aalto.fi

Pierre Massé
UMR STMS
Ircam-CNRS-Sorbonne Université
Paris, France

Vesa Välimäki
Acoustics Lab
Dept. of Signal Processing and Acoustics
Aalto University
Espoo, Finland

Markus Noisternig
UMR STMS
Ircam-CNRS-Sorbonne Université
Paris, France

ABSTRACT

The direction-dependent characteristics of late reverberation have long been assumed to be perceptually isotropic, meaning that the energy of the decay should be perceived equal from every direction. This assumption has been carried into the way reverberation has been approached for spatial sound reproduction. Now that new methods exist to capture the sound field, we need to revisit the way we analyze and render the decaying sound field and more specifically, establish the perceptual threshold of direction-dependent characteristics of late reverberation. Towards this goal, this paper proposes the Energy Decay Deviation (EDD) as an objective measure of the directional decay. Based on the deviation of direction-dependent Energy Decay Curves (EDC) to a mean EDC, the EDD aims to highlight the direction-dependent features characterizing the decay. This paper presents the design considerations of the EDD, discusses its limitations, and shows practical examples of its use.

1. INTRODUCTION

Early in the development of artificial reverberators, it was presumed that no audible direction-dependent characteristics were present in the late reverberant sound field due to our inability to distinguish singular reflections within the decay [1, 2]. When introducing the first digital reverberation algorithm [3], Schroeder suggested that the main requirement for a multichannel reverberator was to produce low correlated signals on multiple loudspeakers. The same assumption was carried on as more sophisticated delay networks where introduced to formalize multichannel reverberation [4, 5]. Contributing to this assumption, a common descriptor of the reverberation is the mixing time (t_{mix}), which is described as the moment where the energy is statistically equal in all regions of a space [6, 7]. Although the exact mathematical definition varies throughout the literature [8], it implies that the energy is expected to remain isotropic once the t_{mix} has been reached.

Reproduction techniques such as Directional Audio Coding (DIRAC) [9] aim to enhance the reproduction of ambisonics recordings through time-frequency analysis by identifying the incident directions of non-diffused sound sources and ensuring they are reproduced with high coherence over a smaller area, while keeping the reverberant part spatially diffused and incoherent. As with previous methods, the key assumption is that the reverberant fields can be reduced to a decorrelated and isotropic signal, usually described as diffused in spatial sound reproduction. Diffuseness is a criteria that describes the spatial correlation in a sound field and was first introduced to assess the required characteristics of reverberant chambers [10] and as such, wasn't originally tied to a reproduction method. Similarly to the mixing time, the formal mathematical definition of the diffusion varies through the literature [10–15]. Nonetheless, once a certain diffusion threshold has been met, the reverberant sound field is here again considered isotropic.

Convolution with a spatial impulse response (SIR) encoded in Ambisonics will naturally preserve direction-dependent decay characteristics as long as the spatial resolution is sufficient enough to prevent high coherence between output signals on the target reproduction system. However, the convolutions required for large loudspeaker array can have prohibitive computational costs in some applications. Sound reproduction methods such as the Spatial Impulse Response Rendering (SIRR) [16] and the Spatial Decomposition Method (SDM) [17] have been used to simplify the reproduction of SIRs and have been shown to perform well in perceived sound quality evaluation. Using the signals from a microphone array, these methods identify the most prominent incident direction(s) of energy within a given short time window, as well as within a frequency band in the case of SIRR. As such, these techniques make no direct assumption on the isotropy of the sound field. However, due to the increased density of echos during the later part of an impulse response, the analysis win-

© Benoit Alary, Pierre Massé, Vesa Välimäki, Markus Noisternig. Licensed under a Creative Commons Attribution 4.0 International License (CC BY 4.0). **Attribution:** Benoit Alary, Pierre Massé, Vesa Välimäki, Markus Noisternig. "Assessing the anisotropic features of spatial impulse responses", 1st EAA Spatial Audio Signal Processing Symposium, Paris, France, 2019.

doi:10.25836/sasp.2019.32

dow may not hold the sufficient resolution for the reproduction of a complex anisotropic decay, meaning they are more suited for isotropic decays. While most multichannel artificial reverberation algorithms do not naturally extend to render direction-dependent decay characteristics, recent work [18] introduced the design principles to build a Directional Feedback Delay Network capable of reproducing anisotropic decay. However, more work remains to analyze the output of such reverberators.

Recent studies have shown our capacity to detect small directional energy variations within the decay of noise signals, suggesting the need for more research to assess the perceptual threshold of direction-dependent characteristics in a decaying sound field [19, 20]. In [21], a perceptual study confirmed that spatial features of the late reverberation contribute to the feeling of envelopment of a listener. However, before a formal perceptual evaluation can be conducted, there is a need for new analysis methods suitable to assess objectively the characteristics of a decaying sound field [22–24].

Towards this goal, this paper proposes an analysis method capable of extracting direction-dependent characteristics within a captured SIR. Based on an existing analysis method, the Energy Decay Curves (EDC) [25], the proposed method consists of calculating the EDC for a set of Directional Impulse Responses (DIRs) [26] and comparing those results with with respect to a mean EDC, calculated from all directions, to obtain the Energy Decay Deviation (EDD). Through this method, the EDD can highlight the anisotropic features of a SIR and as such, is capable of quantifying the direction-dependent characteristics in a decaying sound field, which is an essential step in understanding how these characteristics are perceived by a listener.

Section 2 will cover background information relevant to the proposed method. Section 3 will introduce the EDD method along with possible design choices. In section 4, we will discuss three distinct results obtained with the proposed method. Finally, in the last section, we will discuss future work, research directions and conclude the paper.

2. BACKGROUND

2.1 Energy Decay Curve

The EDC was first introduced as a way to calculate the decay time within a noisy impulse response (IR) [25]. One of the goals of the EDC was to establish the decay time T_{60}. It was later expanded to the time-frequency domain through the Energy Decay Relief (EDR) [27]. The EDC consists of the reverse energy integration from an IR $h(t)$ which can be calculated at a time t with

$$\mathrm{EDC}(t) = \int_t^\infty h^2(\tau)\,\mathrm{d}\tau. \tag{1}$$

2.2 Mixing time

With the mixing time (t_{mix}), we aim to identify the moment where the reverberation transitions from early reflections to diffused late reverberation [6]. For SIRs, the measure of the sound field's diffuseness at different time segments can be used to determine the mixing time. For this purpose, an SIR diffuseness profile was calculated using the spatial covariance method [14] and averaged over the optimal frequency range for a given spherical microphone array. Diffuseness generally begins with low values, due to specular early reflections, and rapidly increases before reaching a relatively stable maximum value (Fig. 1). One way to define t_{mix} is as the moment a stable maximum is reached, and the diffuseness profile shows no more interference from discrete reflections, provided that the maximum is sufficiently high to be considered diffuse (e.g. ≥ 0.9) [28]. Thus the mixing time is calculated as

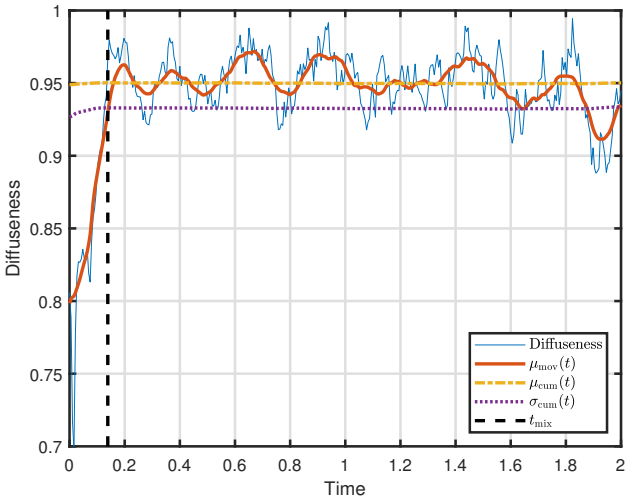


Figure 1. Example mixing time calculation using an SIR captured at the Church of Saint-Eustache in Paris, France, detailed in the results section

$$t_{\mathrm{mix}} = \min(t_d), \tag{2}$$

where t_d are the time values that satisfy the condition

$$\sqrt{\left[\mu_{\mathrm{mov}}(t_d) - \mu_{\mathrm{cum}}(t_d)\right]^2} \leq \sigma_{\mathrm{cum}}(t_d), \tag{3}$$

μ_{mov} is an appropriately-sized moving average of the diffuseness profile, μ_{cum} is its reverse-cumulative average, and σ_{cum} its reverse-cumulative standard deviation.

2.3 Noise floor time

One of the main weaknesses of the EDC integration is the contribution from non-decaying background noise present in a measured IR. Indeed, if the IR contains a long period of noise energy, it can have a significant impact on the EDC calculation and hide some details of the actual energy decay [29]. Several approaches exist to alleviate some of the adverse effects of background noise on the EDC calculation [30–32]. One simple approach is to crop the IR at time t_{noise} where the noise floor starts becoming noticeable in the response.

One method to identify t_{noise} is to first calculate the EDC of the omnidirectional channel of the SIR, then convert it to dB scale and segment the results using an adaptive

doi:10.25836/sasp.2019.32

Ramer-Douglas-Peucker algorithm [28]. The reverse integration of a non-decaying ideal dB-scale noise profile is then fitted by finding the best-matching segment. Finally, t_{noise} is defined as the last point above a specified headroom.

2.4 Beamforming

From the spherical harmonic decomposition of a sound field, we can extract the signals coming from specific incident directions using beamforming methods to obtain directional impulse responses (DIRs). The choice of a beamformer requires a compromise between the main lobe width and its amplitude or energy ratio with respect to the secondary lobes. For a spherical harmonic signal $\mathbf{s}$ at a given discrete time t, the signal y from an incident azimuth ϕ and elevation θ is given by

$$y(t, \phi, \theta) = \mathbf{w}_{\mathrm{PWD}}^{H}\, \mathbf{s}(t), \tag{4}$$

where $\mathbf{w}_{\mathrm{PWD}}$ contains the beamforming weights for the plane wave decomposition (PWD), obtained from

$$\mathbf{w}_{\mathrm{PWD}} = \mathbf{y}(\phi, \theta) \odot \mathbf{d}, \tag{5}$$

which represents the Hadamard product $\odot$ between $\mathbf{y}(\phi, \theta)$, the spherical harmonics vector for a given direction, and a weight vector $\mathbf{d}$ that can be used to change the shape of the beamformer [33, 34].

3. ENERGY DECAY DEVIATION

The main purpose of the EDD is to show the direction-dependent anisotropic characteristics throughout the decay by analyzing the per sample deviations to a mean EDC taken from every direction. Starting from an SIR, the first step is to extract the necessary DIRs for a chosen set of incident directions. From these DIRs, we can then calculate the directional EDCs, mean EDC, and directional EDDs. For analysis purposes, we are interested in a subset of directions that can have a meaningful representation. For instance, we can analyze the lateral plane by fixing the elevation θ to 0 and sampling the azimuth ϕ at fixed intervals.

Using a beamformer, we extract the directional signals from the SIR. As previously mentioned, it is important to bear in mind the shape of the beamformer's lobes when interpreting the results. The width of the main lobe will have a smoothing effect over multiple directions and thus, will reduce the dynamic range of the extracted signals. For the EDD, we chose an hypercardioid type of beamforming for its simplicity and the fact that it can extract signals with a maximum directivity index [33]. For this type of beamforming, the individual values of the $\mathbf{d}$ vector from Eq. 5 are all ones and therefore, the formulation can be simplified to

$$y(t, \phi, \theta) = \mathbf{y}(\phi, \theta)\, \mathbf{s}(t). \tag{6}$$

To prevent the adverse effect of the background noise detailed in Sec.2.3, the DIRs are cropped at time t_{noise}. An alternative method would be to replace the noisy parts of the signal, after t_{noise}, with an artificial noise signal which follows a predicted decaying curve [28].

The early part of the EDD tends to have the larger deviation due to the sparse early reflections. Although the EDD method is capable of showing the direction-dependent contribution of these early-reflections, we recommend cropping the beginning of the EDC before the t_{mix} time. By doing so, we can lower the necessary dynamic range of values used in the analysis of the remaining part of the EDD, which generally contains smaller EDD values. Other well-suited methods exist to analyze the precise direction of arrival (DOA) of early reflections [35]. For the same purpose, we also omit the last few samples of the EDD from the final analysis due to the statistical instability that occurs in very short integration times.

Therefore, we perform the EDC calculations of the extracted DIRs between t_{mix} and t_{noise} through the following reverse integration

$$\mathrm{EDC}(t, \phi, \theta) = \int_{t}^{t_{\mathrm{noise}}} y^2(\tau, \phi, \theta)\, \mathrm{d}\tau, \tag{7}$$

where $t > t_{\mathrm{mix}}$. We then convert these energy curves to the dB scale to perform the analysis on a scale closer to human auditory perception

$$\mathrm{EDC}_{\mathrm{dB}}(t, \phi, \theta) = 10\, \log_{10}(\mathrm{EDC}(t, \phi, \theta)). \tag{8}$$

Calculating the deviation of the EDC in the dB scale is one of the key difference between the proposed method and the method presented earlier in [26], where the same calculations are performed in the linear scale.

The next step to obtain the EDD is to calculate a reference mean $\overline{\mathrm{EDC}}_{\mathrm{dB}}$ for the chosen set of directions. The mean represents an ideal isotropic sound field

$$\overline{\mathrm{EDC}}_{\mathrm{dB}}(t) = \frac{1}{N} \sum_{i=0}^{N} \mathrm{EDC}_{dB}(t, \phi_i, \theta_i). \tag{9}$$

The final EDD is calculated for every directions from the deviation to that mean

$$\mathrm{EDD}(t, \phi, \theta) = \mathrm{EDC}_{\mathrm{dB}}(t, \phi, \theta) - \overline{\mathrm{EDC}}_{\mathrm{dB}}(t). \tag{10}$$

The EDD values represent how much energy remains in the decay when compared to $\overline{\mathrm{EDC}}_{\mathrm{dB}}$. The range of the deviation itself is also an important information in the EDD. Keeping in mind the smoothing caused by the beamformer, it can show narrow deviations which may still be perceived as isotropic. For frequency-dependent analysis, the input signals can be converted to the time-frequency domain before performing the above equations.

4. RESULTS

4.1 Simulated spatial impulse response

To validate the method, an artificial reference signal was generated with controlled direction-dependent characteristics. To construct the signal, we set a target direction-dependant decay time $T_{60}(\phi, \theta)$ for a set of uniformly distributed points around a sphere. We use a shape similar to a

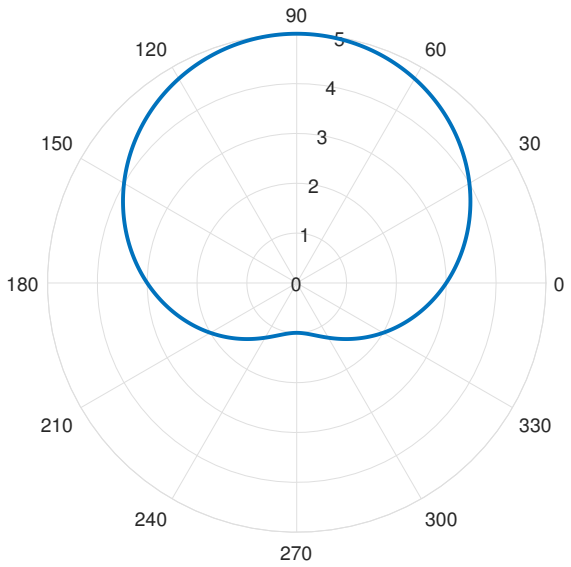


Figure 2. $T_{60}(\phi, 0°)$ of the artificial reference signal to show the distribution on the lateral plane.

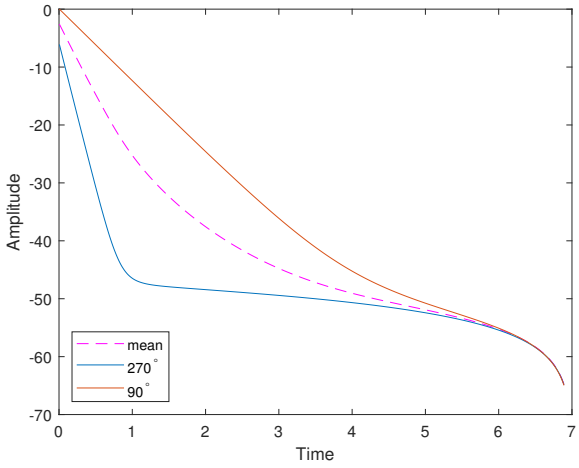


Figure 3. The $\overline{\mathrm{EDC}}_{\mathrm{dB}}$ of the simulated reference signal and the $\mathrm{EDC}_{\mathrm{dB}}(\phi, 0°)$ at $\phi = 90°$ and $\phi = 270°$.

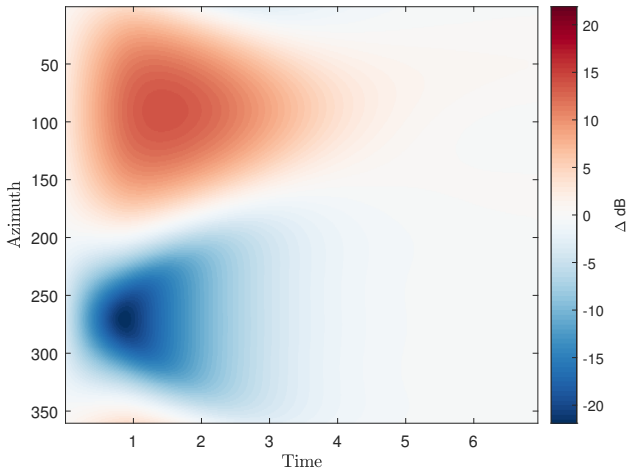


Figure 4. EDD of the reference signal, on the lateral plane.

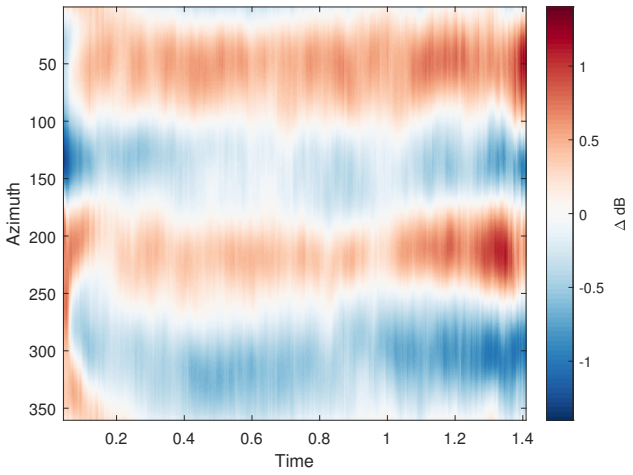


Figure 5. EDD of a long corridor with $t_{\mathrm{mix}} = 42.67$ milliseconds and $t_{\mathrm{noise}} = 1.456$ seconds. The minimum deviation is -1.34 dB and the maximum is 1.21 dB.

cardioid microphone pattern to distribute the $T_{60}(\phi, \theta)$ and create a recognizable shape in the decay (Fig. 2).

For each sampled point (ϕ_i, θ_i), we create an exponentially decaying Gaussian white noise signal that decays at the target rate $T_{60}(\phi_i, \theta_i)$. This signal is then mixed with another noise sequence with an amplitude set to -60 dB which represents the noise floor. These summed noise sequences are then encoded into fourth order ambisonics for each angle pair (ϕ_i, θ_i). The generated SIR signal is then analyzed using the EDD method. In Fig. 3 we can see how two specific $\mathrm{EDC}_{dB}(\phi_i, \theta_i)$ relate to the $\overline{\mathrm{EDC}}_{\mathrm{dB}}$ of the artificial signal. In Fig. 4, we can see the resulting EDDs on the horizontal plane. The red color show points where more energy remains in the decay, therefore showing directions with longer T_{60}, while the blue shows the opposite. The white color represents areas that follows the average. For illustration purposes, since the SIR was generated in controlled settings, the full signal was kept and not cropped as suggested in the method.

4.2 Measurement in a corridor

Next we look at the EDD of a long corridor that is 22 meters long, 2.8 meters wide and 3.2 meters tall. This SIR was recorded with a 32-channel spherical microphone array (Eigenmike®) at fourth order Ambisonics. This corridor is located on the campus of Aalto University in Espoo, Finland. The reverberation was first observed subjectively in that location to have a noticeably longer reverberation time in the direction of the long axis of the corridor. This can be confirmed with the EDD analysis of a captured SIR (Fig. 5). For the capture, the microphone was placed in the center of the room and the loudspeaker was placed 1.5 meters away from the microphone, between the wall and the microphone. In the EDD, we can observe some of the remaining early reflections characteristics before going into a stable direction-dependent energy distribution for the remaining of the decay. The direction-dependent characteristics for this SIR are distributed across 2.55 dB of deviation.

doi:10.25836/sasp.2019.32

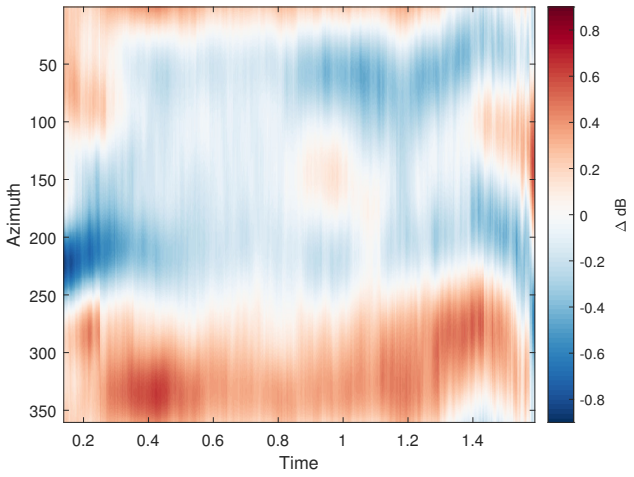


Figure 6. EDD of an SIR captured at the Saint-Eustache church. The $t_{\mathrm{mix}} = 138.67$ milliseconds and $t_{\mathrm{noise}} = 1.637$ seconds. The minimum deviation is -0.88 dB and the maximum is 0.64 dB.

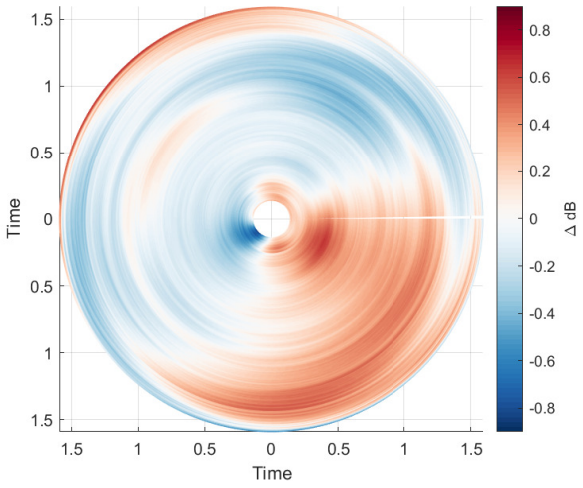


Figure 7. An alternate visualization of the EDD, using the data of Fig. 6, presented on the polar coordinate system to offer a more intuitive representation of directivity.

4.3 Measurement of the Church of Saint-Eustache

Finally, we calculated the EDD from a fourth order Ambisonics SIR captured at the Saint-Eustache church in Paris, France (Fig. 6). In this case, the direction-dependent characteristics are contained within a narrower range, 1.52 dB and it is not known at this point if they are perceptually noticeable. It is important here to reiterate the impact of the beamformer and note it is possible that the spatial smoothing hides larger deviations at narrower directions. Nonetheless, we were able to validate the stability of the EDD by comparing this result between multiple SIRs recorded from the same position (not shown due to space limitations).

We also propose an alternative representation, using a polar coordinate system, in which the radius represents time and the angles correspond to the incident direction of sound (Fig. 7). This form can show the characteristics more intuitively in applications where a lower resolution in the early decay is an acceptable compromise.

5. CONCLUSION

In conclusion, we introduced a formal method to analyze and assess the anisotropic features of an SIR. Based on the EDC, a popular analysis method for IRs, the proposed EDD measure can highlight direction-dependent characteristics in a decaying sound field. As such, it can serve as an analysis tool to study the behavior of late reverberation in more depth. This method benefits from using higher order ambisonics, to minimize the spatial smoothing caused by a beamformer. Future work includes analyzing different SIRs and conducting in depth perceptual studies to connect the directional characteristics of reverberation to the human auditory perception, as well as validating the use of denoising techniques to overcome the necessity of cropping the signal before the analysis.

6. ACKNOWLEDGEMENTS

Part of this work was conducted during Benoit Alary's research visit in October–December 2018, which was hosted by the UMR STMS (IRCAM-CNRS-Sorbonne Université) and was funded by the Foundation for Aalto University Science and Technology and by the Academy of Finland (ICHO project, Aalto University project no. 13296390).

7. REFERENCES

[1] A. D. Pierce, "Concept of a directional spectral energy density in room acoustics," *J. Acoust. Soc. Am.*, vol. 56, pp. 1304–1305, Oct. 1974.

[2] V. Välimäki, J. D. Parker, L. Savioja, J. O. Smith, and J. S. Abel, "Fifty years of artificial reverberation," *IEEE Trans. Audio Speech Lang. Process.*, vol. 20, pp. 1421–1448, Jul. 2012.

[3] M. R. Schroeder and B. F. Logan, "'Colorless' artificial reverberation," *J. Audio Eng. Soc.*, vol. 9, pp. 192–197, Jul. 1961.

[4] M. A. Gerzon, "Synthetic stereo reverberation, part I and II," *Studio Sound*, vol. 13(I), 14(II), pp. 632–635(I), 24–28(II), Dec. 1971(I), Jan. 1972(II).

[5] J. Stautner and M. Puckette, "Designing multi-channel reverberators," *Computer Music J.*, vol. 6, pp. 52–65, Spring 1982.

[6] J.-M. Jot, L. Cerveau, and O. Warusfel, "Analysis and synthesis of room reverberation based on a statistical time-frequency model," in *Proc. AES 103rd Conv.*, (New York, USA), Sep. 1997.

[7] B. A. Blesser, "An interdisciplinary synthesis of reverberation viewpoints," *J. Audio Eng. Soc*, vol. 49, pp. 867–903, Oct. 2001.

[8] S. Schlecht and E. Habets, "Feedback delay networks: Echo density and mixing time," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 25, pp. 374–383, Feb. 2017.

[9] V. Pulkki, "Spatial sound reproduction with directional audio coding," *J. Audio Eng. Soc.*, vol. 55, pp. 503–516, Jun. 2007.

[10] C. G. Balachandran and D. W. Robinson, "Diffusion of the decaying sound field," *Acta Acust.*, vol. 19, pp. 245–257, Jan. 1967.

[11] G. Bart, "Spatial cross-correlation in anisotropic sound fields," *Acta Acust.*, vol. 28, pp. 45–49, Jan. 1973.

[12] H. Cox, "Spatial correlation in arbitrary noise fields with application to ambient sea noise," *J. Acoust. Soc. Am.*, vol. 54, pp. 1289–1301, Nov. 1973.

[13] R. Talham, "Noise correlation functions for anisotropic noise fields," *J. Acoust. Soc. Am.*, vol. 69, pp. 213–215, Jan. 1981.

[14] N. Epain and C. T. Jin, "Spherical harmonic signal covariance and sound field diffuseness," *CoRR*, vol. abs/1607.00211, Jul. 2016.

[15] C.-H. Jeong, M. Nolan, and J. Balint, "Difficulties in comparing diffuse sound field measures and data/code sharing for future collaboration," in *Proc. Euronoise 2018*, (Heraklion, Crete), pp. 2005–2010, May 2018.

[16] J. Merimaa and V. Pulkki, "Spatial impulse response rendering I: Analysis and synthesis," *J. Audio Eng. Soc.*, vol. 53, pp. 1115–1127, Dec. 2005.

[17] S. Tervo, J. Ptynen, A. Kuusinen, and T. Lokki, "Spatial decomposition method for room impulse responses," *J. Audio Eng. Soc.*, vol. 61, pp. 17–28, Jan. 2013.

[18] B. Alary, A. Politis, S. J. Schlecht, and V. Välimäki, "Directional feedback delay network," *J. Audio Eng. Soc.*, in press 2019.

[19] D. Romblom, C. Guastavino, and P. Depalle, "Perceptual thresholds for non-ideal diffuse field reverberation," *J. Acoust. Soc. Am.*, vol. 140, pp. 3908–3916, Nov. 2016.

[20] P. Luizard, B. F. G. Katz, and C. Guastavino, "Perceptual thresholds for realistic double-slope decay reverberation in large coupled spaces," *J. Acoust. Soc. Am.*, vol. 137, pp. 75–84, Jan. 2015.

[21] W. Lachenmayr, A. Haapaniemi, and T. Lokki, "Direction of late reverberation and envelopment in two reproduced Berlin concert halls," in *Proc. AES 140th Conv.*, (Paris, France), Jun. 2016.

[22] M. Nolan, E. Fernandez-Grande, J. Brunskog, and C.-H. Jeong, "A wavenumber approach to quantifying the isotropy of the sound field in reverberant spaces," *J. Acoust. Soc. Am.*, vol. 143, pp. 2514–2526, Apr. 2018.

[23] T. Sakuma and K. Eda, "Energy decay analysis of non-diffuse sound fields in rectangular rooms," *Proc. of Meetings on Acoustics*, vol. 19, p. 015138, Jun. 2013.

[24] S. Oksanen, J. Parker, A. Politis, and V. Välimäki, "A directional diffuse reverberation model for excavated tunnels in rock," in *Proc. IEEE ICASSP-13*, (Vancouver, Canada), pp. 644–648, May 2013.

[25] M. R. Schroeder, "New method of measuring reverberation time," *J. Acoust. Soc. Am.*, vol. 37, pp. 409–412, Mar. 1965.

[26] M. Berzborn and M. Vorländer, "Investigations on the directional energy decay curves in reverberation rooms," in *Proc. Euronoise 2018*, (Heraklion, Crete, Greece), pp. 2005–2010, May 2018.

[27] J.-M. Jot, "An analysis/synthesis approach to real-time artificial reverberation," in *Proc. IEEE ICASSP-92*, vol. 2, (San Francisco, CA), pp. 221–224, Mar. 1992.

[28] P. Massé, T. Carpentier, O. Warusfel, and M. Noisternig, "Refinement and implementation of a robust directional room impulse response denoising process, including applications to highly varied measurement databases," in *Proc. ICSV26*, (Montréal, Canada), Jul. 2019.

[29] M. Karjalainen, P. Antsalo, A. Mäkivirta, T. Peltonen, and V. Välimäki, "Estimation of modal decay parameters from noisy response measurements," *J. Audio Eng. Soc.*, vol. 50, pp. 867–878, Nov. 2002.

[30] D. R. Morgan, "A parametric error analysis of the backward integration method for reverberation time estimation," *J. Acoust. Soc. Am.*, vol. 101, pp. 2686–2693, May 1997.

[31] L. Faiget, C. Legros, and R. Ruiz, "Optimization of the impulse response length: Application to noisy and highly reverberant rooms," *J. Audio Eng. Soc.*, vol. 46, pp. 741–750, Sep. 1998.

[32] Y. Hirata, "A method of eliminating noise in power responses," *J. Sound Vibr.*, vol. 82, pp. 593–595, Jun. 1982.

[33] B. Rafaely, *Fundamentals of Spherical Array Processing*. Springer, 2015.

[34] L. McCormack, S. Delikaris-Manias, and V. Pulkki, "Parametric acoustic camera for real-time sound capture, analysis and tracking," in *Proc. DAFx-17*, (Edinburgh, UK), pp. 412–419, Sep. 2017.

[35] S. Tervo and A. Politis, "Direction of arrival estimation of reflections from room impulse responses using a spherical microphone array," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 23, pp. 1539–1551, Oct. 2015.

 doi:10.25836/sasp.2019.32

HOW WEARING HEADGEAR AFFECTS MEASURED HEAD-RELATED TRANSFER FUNCTIONS

Christoph Pörschmann[1] **Johannes M. Arend**[1,2] **Raphael Gillioz**[3]
[1] TH Köln, Institute of Communications Engineering, Cologne, Germany
[2] TU Berlin, Audio Communication Group, Berlin, Germany
[3] TH Köln, Institute of Media and Imaging Technology, Cologne, Germany
Christoph.Poerschmann@th-koeln.de

ABSTRACT

We present spherical high-density measurement data of head-related transfer functions (HRTFs) and analyze the influence of wearing headgear during the measurements. For this we captured datasets from a Neumann KU100 and a HEAD acoustics HMS II.3 dummy head either equipped with a bicycle helmet, a baseball cap, an Oculus Rift head-mounted display, or AKG K1000 headphones. We investigate the influence of different types of headgear in terms of their spectrum and their binaural cues and compare the results to reference measurements of the dummy heads without headgear. Generally, the results show that differences to the reference vary significantly depending on the type of the headgear. The spectral differences to the reference are maximal for the AKG K1000 and smallest for the Oculus Rift and the baseball cap. Analyzing the influence of the incidence directions on the spectral differences we found the strongest deviations for the Oculus Rift and the baseball cap for contralateral sound incidence. For the bicycle helmet, the contralateral directions were also most affected, but shifted upwards in elevation. Finally, for the AKG K1000, which generally has the highest impact on the spectrum of the HRTFs, we observed maximal deviations for sound incidence from behind. Regarding the interaural time differences (ITDs) and interaural level differences (ILDs) the analysis again revealed the highest influences for the AKG K1000. While for the Oculus Rift the ITDs and ILDs were mainly affected for frontal directions, we observed only a very weak influence of the bicycle helmet and the baseball cap. The HRTF sets are available in the SOFA format under a Creative Commons CC BY-SA 4.0 license.

1. INTRODUCTION

The spatial representation of sound sources is an essential element of virtual acoustic environments (VAEs). When determining the sound incidence direction, the human auditory system evaluates monaural and binaural cues, which are caused by the shape of the pinna and the head. While spectral information is the most important cue for elevation of a sound source, we use differences between the signals reaching the left and the right ear for lateral localization. These binaural differences manifest in interaural time differences (ITDs) and interaural level differences (ILDs). In many headphone-based VAEs, head-related transfer functions (HRTFs) are used to describe the sound incidence from a source to the left and right ear, thus integrating both monaural and binaural cues [1]. Various researchers investigated the perceptual influences of individual measurements of HRTFs (e.g. [2, 3]) and found improvements regarding externalization, localization as well as a reduction of front-back confusions when using individualized datasets. Furthermore, specific properties, like for example the torso [4], and probably even headgear [5–7] influence the HRTFs and thus as well localization and other perceptual attributes.

Generally speaking, apart from individualization and head-above-torso movements, in many real-life situations spatial cues are modified by headgear, for example by wearing a baseball cap, a bicycle helmet, or a head-mounted display (HMD), which nowadays is often used in VR applications. However, often a good localization performance is important when wearing such items, e.g. in order to determine approaching vehicles when cycling. Furthermore, when performing psychoacoustic experiments in mixed-reality applications using HMDs, the influence of the HMD on the HRTFs must be considered. Effects of an HTC Vive HMD on localization performance have already been analyzed by Ahrens et al [8]. The authors performed a loudspeaker-based localization experiment presenting conditions with and without an HMD, which showed significant differences in localization due to the HMD. To analyze the influence of headgear for varying directions of incidence, measurements of HRTFs on a dense spherical sampling grid are required. However, HRTF measurements of a dummy head with various headgear are still rare, and to our knowledge only one dataset measured for an HTC Vice on a sparse grid with 64 positions is freely accessible [8].

To accurately analyze these influences for all directions of incidence, measurements of HRTFs on a dense

© Christoph Pörschmann, Johannes M. Arend, Raphael Gillioz. Licensed under a Creative Commons Attribution 4.0 International License (CC BY 4.0). **Attribution:** Christoph Pörschmann, Johannes M. Arend, Raphael Gillioz. "How Wearing Headgear Affects Measured Head-Related Transfer Functions", 1st EAA Spatial Audio Signal Processing Symposium, Paris, France, 2019.

doi:10.25836/sasp.2019.27



(a) (b) (c) (d)

Figure 1: Measurement of the HRTF set in the anechoic chamber at TH Köln for the baseball cap on the HEAD acoustics HMS II.3 (a), the Oculus Rift on the Neumann KU100 (b), the Uvex bicycle helmet on the Neumann KU100 (c), and the AKG K1000 on the HEAD acoustics HMS II.3 (d).

grid need to be performed. Several authors suggested to describe complete sets of HRTFs in spherical harmonics (SH) domain [9, 10]. Here, the HRTF set, measured on a spherical grid, is decomposed into spherical base functions of different spatial orders N, where higher orders correspond to a higher spatial resolution. To completely consider these properties and to avoid spatial aliasing, an order $N \geq kr$ with $k = \omega/c$, and r being the head radius is required [11, 12]. Assuming $r = 8.75\,\text{cm}$ as the average human head radius [13] and $c = 343\,\text{m/s}$ leads to $N = 32$ requiring at least 1089 measured directions. In this case no spatial aliasing occurs for frequencies up to 20 kHz. However, only a few of the datasets mentioned above were measured spherically with an adequate density, none of them investigating the influence of headgear.

This work presents high-density measurement data of HRTF sets from a Neumann KU100 and a HEAD acoustics HMS II.3 dummy head, either equipped with a bicycle helmet, a baseball cap, an Oculus Rift HMD, or a set of extra-aural AKG K1000 headphones captured on a full spherical Lebedev grid with 2702 points. We analyze the datasets in terms of their spectrum and their binaural cues and compare the results to reference measurements of the dummy heads without headgear.

2. HRTF MEASUREMENTS

We performed HRTF measurements of a Neumann KU100 dummy head and a Head acoustics HMS II.3 for four different types of headgear: a baseball cap (Flexfit Snapback), a bicycle helmet (Uvex Cobra RS), an HMD (Oculus Rift, without any kind of headphones) and AKG K1000 extraaural headphones. As loudspeaker we used a Genelec 1029A, which has a flat on-axis frequency response from 50 Hz to 20 kHz (±3 dB). The HRTFs were measured on a Lebedev full spherical grid with 2702 points. We applied the VariSphear measurement system [14] for precise positioning of the dummy head at the spatial sampling positions and for capturing the HRTFs. Besides the motor control and impulse response capture modules, the software provides an automatic error detection which checks every measured impulse response for noticeable variations compared to the previous measurement. This ensures a validity of all obtained impulse responses. The excitation signal for all measurements was an emphasized sine sweep with +20 dB low shelf at 100 Hz (2^{18} samples at 48 kHz sampling rate, length 5.5 s). An RME Babyface audio interface served as AD / DA converter and microphone preamp. For further details on the measurement set-up and procedure please refer to [15, 16].

The measurements were carried out in the anechoic chamber at TH Köln. The room has dimensions of $4.5\,\text{m} \times 11.7\,\text{m} \times 2.3\,\text{m}$ and a lower cut-off frequency of about 200 Hz. Fig. 1 shows the Neumann KU100 and the Head acoustics HMS II.3 with the different types of headgear mounted on the VariSphear device. The height of the loudspeaker and of the dummy head was at 1.25 m and the acoustic center of the loudspeaker was always set to the ear level of the dummy head. All sets of HRTFs were captured at 2 m distance. For all setups, exact alignment of the head was checked for various sampling positions. The distance between the loudspeaker and the entrance of the dummy head's ear canal was for each measurement accurately determined with a laser distance meter. Additionally, we used a Microtech Gefell M296S microphone positioned at the acoustic center of the dummy head to measure omnidirectional impulse responses which we used for the magnitude and phase compensation of the loudspeaker.

In a subsequent postprocessing the raw measurement data were first carefully truncated and windowed. Then we compensated the influence of the loudspeaker by inverse FIR filtering with the measured omnidirectional impulse response. The final length of each HRIR is 128 samples at a sampling rate of 48 kHz. The postprocessing is based on the implementation and description from [15, 16]. A fur-

doi:10.25836/sasp.2019.27

ther step of the postprocessing eliminates low-frequency artefacts resulting from the frequency response of the relatively small loudspeaker, which fails to reproduce low frequencies at adequate sound pressure levels. Additionally this step removes the influence of room modes and reflections arising from the sound field of the anechoic chamber below its cut-off frequency. A well-suited approach is to replace the low-frequency range of the HRTFs by an analytic expression assuming that for low frequencies (e.g. below 200 Hz), pinna and ear canal hardly affect the HRTF and even the spherical shape of the head only has minor influence on the sound field. In this study we use a low-frequency extension in the frequency domain according to [17] and apply a linear cross-fade between the low-frequency component and the raw HRTFs in a crossover frequency range from 200 Hz – 400 Hz. The level is calculated from the mean absolute values, while the phase is linearly extrapolated in the crossover frequency range.

Finally we transformed the dataset to the spherical harmonics (SH) domain and stored it in form of SH coefficients [9, 10]. This allows for a calculation of the HRTFs at any direction by means of the inverse SH transform.

Even though the measurements were conducted with great care, there are several influencing factors which should be considered. The influence of the robot arm of the VariSphear system on sound radiation to the ear is hard to quantify and depends very much on frequency and incidence direction. Please refer to [15, 16] for a more detailed analysis of these influences. However, for our comparisons of the measured datasets to a reference without headgear these influences are quite irrelevant as they occur in the same way in all measured HRTF sets. In the context of our study, differences which are induced by small variations, e.g. by non exact placement of the loudspeaker and the dummy head are more important. We minimized these inaccuracies by exactly positioning and calibrating the device before each measurement session.

3. RESULTS

3.1 Spectrum

Fig. 2 shows the magnitude responses for frontal and contralateral sound incidence both for the reference and the different headgear. Below 1 kHz the differences are quite small because the wavelength is larger than the geometric structures of the headgear. Generally, the results show that the deviations to the reference are minor for the baseball cap and the Oculus Rift. However, for the baseball cap we observed differences of more than 10 dB at frequencies between 5 kHz and 8 kHz for contalateral sound incidence. The bicycle helmet has nearly no influence for frontal sound incidence, but contralaterally already below 3 kHz strong comb-filter effects become apparent. For all headgear tested, the differences to the reference are maximal for the AKG K1000. Even though the headphones do not directly cover the ears, they strongly influence sound incidence. Accodingly, for frontal sound incidence already at 2 kHz the magnitude response is reduced by 10 dB.

To analyze the spectral deviations to a reference set, we calculated the averaged values across all 2702 measured directions $\Omega\{(\phi_1, \theta_1), \ldots, (\phi_T, \theta_T)\}$:

$$\Delta G_f(\omega) = \frac{1}{n_\Omega} \sum_{\Omega=1}^{n_\Omega} |20lg \frac{|\mathit{HRTF}_{\mathit{REF}}(\omega, \Omega)|}{|\mathit{HRTF}_{\mathit{TEST}}(\omega, \Omega)|}|, \quad (1)$$

with ω describing the temporal frequency, $\mathit{HRTF}_{\mathit{REF}}$ the reference HRTF set and $\mathit{HRTF}_{\mathit{TEST}}$ the HRTF set of the tested headgear.

Fig. 3 illustrates the frequency-dependent spectral differences $\Delta G_f(\omega)$ between the reference without headgear and the different headgear both for the KU100 and the HMS II.3. Below 1 kHz the deviations are mostly in the range of 1 dB or below. Only for the AKG K1000 on the HMS II.3 dummy head they exceed 2 dB. Generally, we observed the lowest differences to the reference for the baseball cap and the Oculus Rift which are in the range of 2 dB for frequencies up to 10 kHz. The deviations are higher for the bicycle helmet, reaching 4 dB already at frequencies below 10 kHz. Finally the highest differences of all tested headgear occur for the AKG K1000. Already at 2 kHz the values exceed 4 dB and reach 6 dB at about 10 kHz.

In a next step we analyzed the spatial distribution of the differences and calculated the directional deviation across all frequencies as

$$\Delta G_{sp}(\Omega) = \frac{1}{n_\omega} \sum_{\omega=1}^{n_\omega} | \, 20lg \frac{| \, \mathit{HRTF}_{\mathit{REF}}(\omega, \Omega) \, |}{| \, \mathit{HRTF}_{\mathit{TEST}}(\omega, \Omega) \, |} \, |, \quad (2)$$

In this case the sampling grid Ω_t was full spherical, in steps of 1° for azimuth and elevation. The results are very similar for both tested dummy heads. Fig. 4 shows the results for the Neumann KU100. For the baseball cap (a) and for the Oculus Rift (c) the spectral differences are mainly located at contralateral directions. Here, the sound incidence which is dominated by diffraction around the the head is strongly affected. For the different types of headgear we found maximal deviations $\Delta G_{sp,max}$ in a range of $8\,dB$ to $10\,dB$. For the baseball cap the maximal deviations are located at $\phi = 272^\circ$ and $\theta = 13^\circ$ and for the Oculus Rift at $\phi = 260^\circ$ and $\theta = 10^\circ$. As the bicycle helmet mainly covers the top of the head we observed maximal spectral differences shifted upwards at $\phi = 273^\circ$ and $\theta = 59^\circ$. For the AKG K1000 the plot shows distinct spectral differences spread over the entire angular range with large areas of high spectral deviations. We observed the highest deviations at $\phi = 220^\circ$ and $\theta = -9^\circ$.

3.2 Binaural cues

Next we compared the ILDs and ITDs of the different headgear to the reference without headgear. For this purpose, we extracted HRTFs in the horizontal plane ($\theta = 0^\circ$) with an angular spacing of $\phi = 1^\circ$ from the datasets. The broadband ILDs were then calculated as the ratio between the energy of the left and right ear HRIR. The ITDs were calculated from the HRIRs by applying the threshold onset method with ten-times oversampling for more precise onset detection.

doi:10.25836/sasp.2019.27

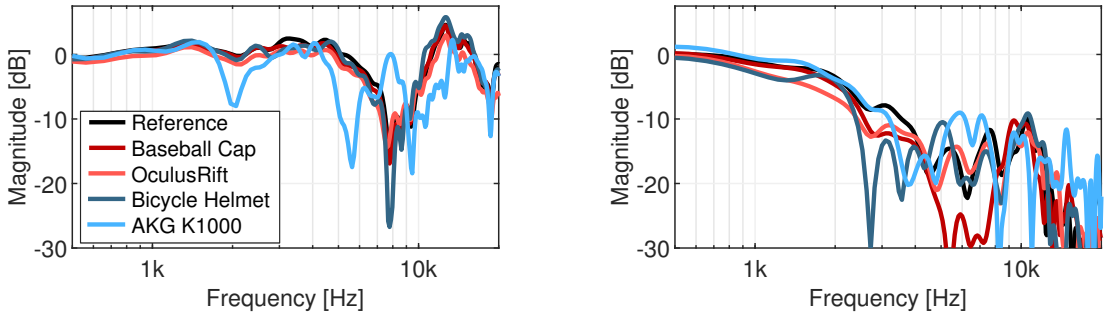


Figure 2: Left ear magnitude, extracted from the sets of the reference (black) and the different headgear on a Neumann KU100. In red the results for the baseball cap and the bicycle helmet are shown, in blue the results for the Oculus Rift and for the AKG K1000. (a) Front direction ($\phi = 0°, \theta = 0°$). (b) Contralateral direction ($\phi = 270°, \theta = 0°$).

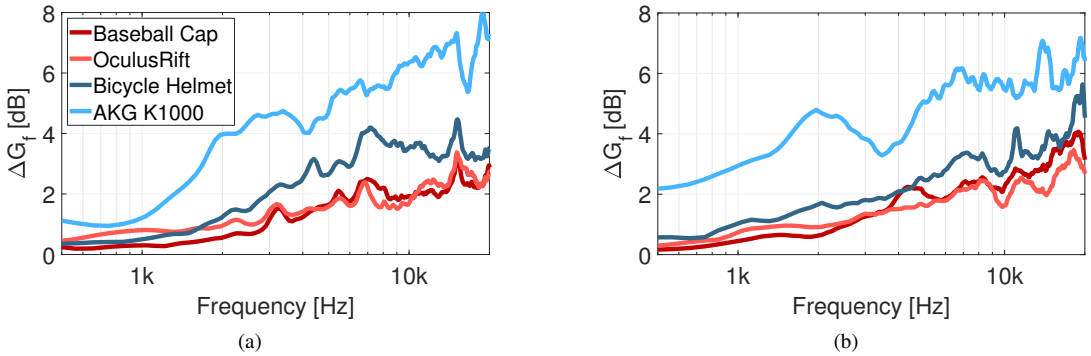


Figure 3: Spectral differences $\Delta G_f(\omega)$ in dB (left ear) between reference HRTF set and the different headgear. In red the results for the baseball cap and the bicycle helmet are shown, in blue the results for the Oculus Rift and for the AKG K1000. (a) Spectral differences for Neumann KU100, (b) for HEAD acoustics HMS II.3.

Fig. 5 shows the calculated ILDs and ITDs of the reference HRTF set and the different headgear. As depicted in Fig. 5 (a,c), for the baseball cap and the bicycle helmet the ILDs are similar to the reference without headgear. Only marginal deviations can be observed here, mainly at lateral directions. For the Oculus Rift stronger deviations occur and for the AKG K1000 the ILDs are altered by up to 4 dB. As can be seen in Fig. 5 (b,d), the ITDs of the measured HRTF sets are generally in good agreement with the ITDs of the reference. While only small deviations exist for the baseball cap, the bicycle helmet and the HMD, the ITDs are significantly reduced laterally for the AKG K1000 measured on the Neumann KU 100.

4. DISCUSSION

In the previous section we have shown that headgear significantly affects measured HRTFs. Now we discuss how this relates to some other factors influencing HRTF measurements. In this context Brinkmann et al. [4] analyzed deviations due to different head-above-torso rotations. The results showed direction-dependent spectral deviations comparable to the ones of the baseball cap or the Oculus Rift. Furthermore, the authors performed a listening experiment showing the audibility of these deviations.

In the context of HRTF measurements, spectrum, temporal structure and interaural differences are as well affected by spatial upsampling. For sparse HRTF sets, which are often used for individual measurements, a subsequent spatial upsampling needs to be performed, e.g. by applying a spatial Fourier transform of the data in the spherical harmonics domain and resampling the HRTFs by an inverse transformation on a dense sampling grid [9, 10]. However, this results in spatial aliasing and truncation errors. The influence of these so-called sparsity errors on the upsampling of HRTFs has for example been analyzed in [18, 19]. Depending on the spatial resolution of the sparse HRTF set, the contributions of the baseball cap, the Oculus Rift and even the bicycle helmet are in the same range or even lower than the sparsity errors. For example, for spatial upsampling by spherical harmonic interpolation at a spatial order of $N = 13$, resulting in 266 measurements on a Lebedev grid, the mean spectral deviations are still above 4 dB at frequencies of 10 kHz [19]. Even when performing an improved interpolation method according to Pörschmann et al. [19], the averaged spectral deviations on this grid are still in the same range as for the baseball cap or the Oculus Rift. Comparing the results of our study to investigations on spatial upsampling of sparse HRTF sets reveals, that a specific focus must be put on the spatial upsampling when

doi:10.25836/sasp.2019.27

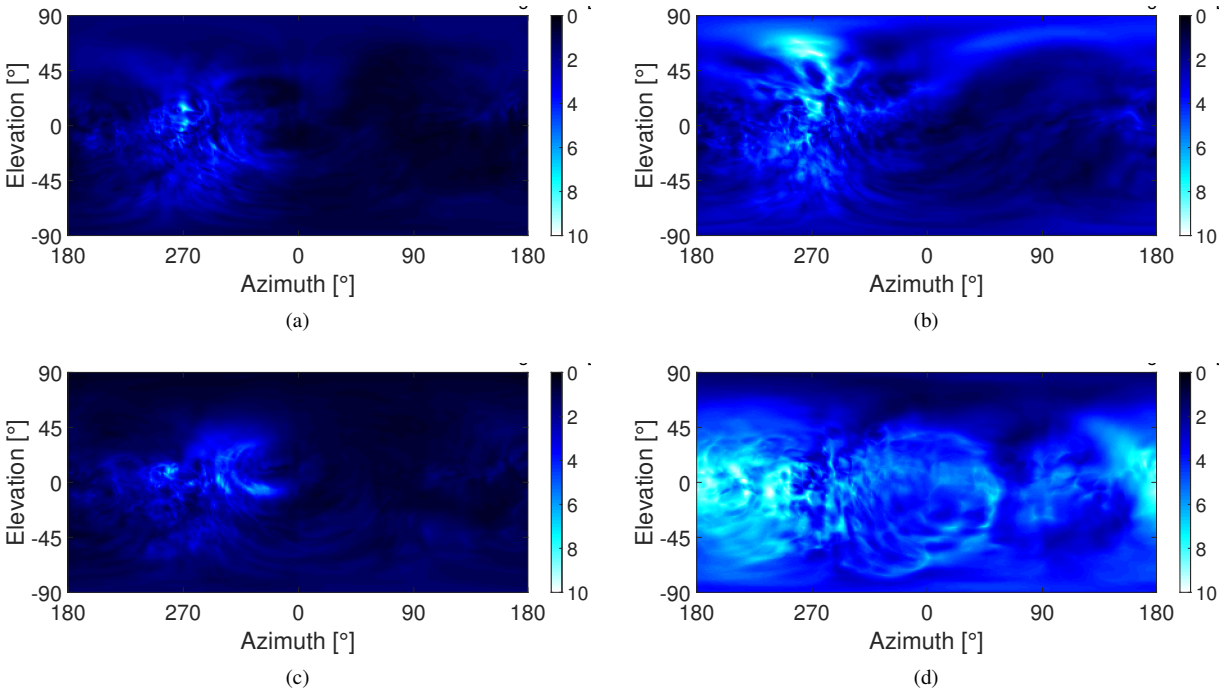


Figure 4: Spectral differences $\Delta G_{sp}(\Omega_t)$ per sampling point and $f \leq 10\,\text{kHz}$ for the different types of headgear and the KU100: baseball cap (a) bicycle helmet (b) , Oculus Rift (c) and the AKG K1000 (d).

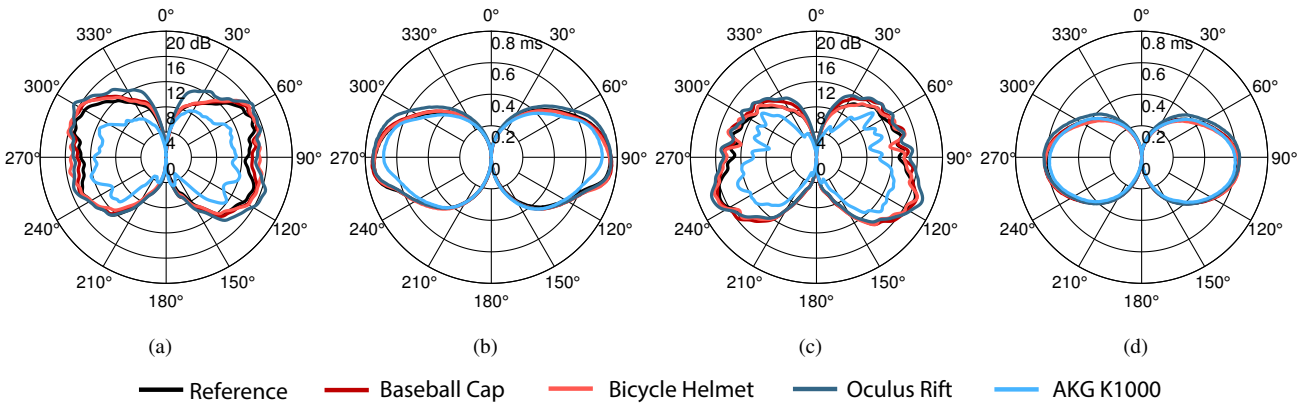


Figure 5: ILDs (a,c), and ITDs (b,d) in the horizontal plane for the reference HRTF set (black) and for the different headgear. The angle represents the azimuth ϕ of the sound incidence. The radius describes the magnitude of the level differences (in dB) or time differences (in ms). Results for the Neumann KU100 (a,b). Results for the HEAD acoustics HMS II.3 (c,d).

using sparse sets of low spatial order, more than on the influence of headgear.

5. CONCLUSION

In this paper we have described a series of measurements of spherical HRTF sets with two different dummy heads equipped with various headgear. We analyzed their influence on the spectrum and on the binaural cues. The results show that differences to the reference without headgear vary significantly depending on the type of the headgear. Regarding the ITDs and ILDs, the analysis revealed the highest influences for the AKG K1000. While for the Oculus Rift HMD, the ITDs and ILDs are affected strongest for frontal directions, generally only a very weak influence of the bicycle helmet and the baseball cap on ITDs and ILDs was observed. This suggests that localization in the horizontal plane is hardly affected by the headgear. The spectral differences to the reference are maximal for the AKG K1000, lowest for the Oculus Rift and the baseball cap. Furthermore, we analyzed for which incidence directions the spectrum is influenced most by the headgear. For the Oculus Rift and the baseball cap, the strongest deviations were found for contralateral sound incidence. For the bicycle helmet, the directions mostly affected are as well contralateral, but slightly shifted upwards in elevation. Finally,

doi:10.25836/sasp.2019.27

the AKG K1000 headphones generally have the highest impact on the measured HRTFs, which becomes maximal for sound incidence from behind.

The results of this study are relevant for applications where headgear is worn and localization or other aspects of spatial hearing are considered. This could be the case in mixed-reality applications where natural sound sources are presented while the listener is wearing an HMD, or when investigating localization performance in certain situations, e.g. in sports activities where headgear is used. Of course, our findings need to be verified for individually measured HRTF sets and be validated in a subsequent perceptual evaluation. However, it was the primary intention of this study to provide freely available HRTF sets which are well-suited for auralization purposes and which allow to further investigate the influence of headgear on auditory perception. The HRTF sets are available in the SOFA format under a Creative Commons CC BY-SA 4.0 license and can be downloaded at: http://audiogroup.web.th-koeln.de/headgear.html. The research presented in this paper has been funded by the German Federal Ministry of Education and Research, support Code: BMBF 03FH014IX5-NarDasS.

6. REFERENCES

[1] J. Blauert, *Spatial Hearing - The Psychophysics of Human Sound Localization*. Cambridge, MA: MIT Press, revised ed., 1996.

[2] E. M. Wenzel and S. H. Foster, "Perceptual consequences of interpolating head-related transfer functions during spatial synthesis," in *Proc. of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pp. 102–105, 1993.

[3] D. R. Begault, E. M. Wenzel, and M. R. Anderson, "Direct comparison of the impact of head tracking, reverberation, and individualized head-related transfer functions on the spatial perception of a virtual speech source.," *Journal of the Audio Engineering Society*, vol. 49, no. 10, pp. 904–916, 2001.

[4] F. Brinkmann, R. Roden, A. Lindau, and S. Weinzierl, "Audibility and Interpolation of Head-Above-Torso Orientation in Binaural Technology," *IEEE Journal on Selected Topics in Signal Processing*, vol. 9, no. 5, pp. 931–942, 2015.

[5] G. Wersényi and A. Illényi, "Differences in Dummy-head HRTFs caused by the Acoustical Environment Near the Head," *Electronic Journal of Technical Acoustics*, vol. 1, pp. 1—-15, 2005.

[6] G. Wersényi and J. Répás, "Comparison of HRTFs from a Dummy-Head Equipped with Hair, Cap and Glasses in a Virtual Audio Listening Task over Equalized Headphones," in *Proc. of the 142nd AES Convention, Berlin*, 2017.

[7] R. Gupta, R. Ranjan, J. He, and W.-s. Gan, "Investigation of effect of VR/AR headgear on Head related transfer functions for natural listening," in *Proc. of the AES Conference on Audio for Virtual and Augmendet Reality*, no. August, 2018.

[8] A. Ahrens, K. D. Lund, M. Marschall, T. Dau, and H. S. Group, "Sound source localization with varying amount of visual information in virtual reality," *PLoS ONE*, vol. 14, pp. 1–19, 2019.

[9] E. G. Williams, *Fourier Acoustics - Sound Radiation and Nearfield Acoustical Holography*. London, UK: Academic Press, 1999.

[10] B. Rafaely, *Fundamentals of Spherical Array Processing*. Berlin Heidelberg: Springer-Verlag, 2015.

[11] B. Rafaely, "Analysis and Design of Spherical Microphone Arrays," *IEEE Transaction on Speech and Audio Processing*, vol. 13, no. 1, pp. 135–143, 2005.

[12] B. Bernschütz, A. Vázquez Giner, C. Pörschmann, and J. M. Arend, "Binaural reproduction of plane waves with reduced modal order," *Acta Acustica united with Acustica*, vol. 100, no. 5, pp. 972–983, 2014.

[13] R. V. L. Hartley and T. C. Fry, "The Binaural Location of Pure Tones," *Physical Review*, vol. 18, no. 6, pp. 431–442, 1921.

[14] B. Bernschütz, C. Pörschmann, S. Spors, and S. Weinzierl, "Entwurf und Aufbau eines variablen sphärischen Mikrofonarrays für Forschungsanwendungen in Raumakustik und Virtual Audio," in *Proc. of the 36th DAGA*, pp. 717–718, 2010.

[15] B. Bernschütz, "A Spherical Far Field HRIR / HRTF Compilation of the Neumann KU 100," in *Proc. of the 39th DAGA*, pp. 592–595, 2013.

[16] J. M. Arend, A. Neidhardt, C. Pörschmann, P. Christoph, and C. Pörschmann, "Measurement and Perceptual Evaluation of a Spherical Near-Field HRTF Set," in *Proc. of the 29th Tonmeistertagung - VDT International Convention*, pp. 52–55, 2016.

[17] B. Xie, "On the low frequency characteristics of head-related transfer function," *Chinese J. Acoust.*, vol. 28, pp. 1–13, 2009.

[18] F. Brinkmann and S. Weinzierl, "Comparison of head-related transfer functions pre-processing techniques for spherical harmonics decomposition," in *Proc. of the AES International Conference on Audio for Virtual and Augmented Reality*, pp. 1–10, 2018.

[19] C. Pörschmann, J. M. Arend, and F. Brinkmann, "Directional Equalization of Sparse Head-Related Transfer Function Sets for Spatial Upsampling," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 6, pp. 1060 – 1071, 2019.

 doi:10.25836/sasp.2019.27

INFLUENCE OF VISION ON SHORT-TERM SOUND LOCALIZATION TRAINING WITH NON-INDIVIDUALIZED HRTF

Tifanie Bouchara[1] **Tristan-Gaël Bara**[1 2]
Pierre-Louis Weiss[2] **Alma Guilbert**[2]
[1] CEDRIC (EA4626), CNAM, HeSam Université ,75003, Paris, France
[2] VAC Laboratory (EA 7326), Université Paris Descartes, 92774 Boulogne-Billancourt, Paris, France
tifanie.bouchara@cnam.fr, alma.guilbert@parisdescartes.fr

ABSTRACT

Previous studies have demonstrated that it is possible for humans to adapt to new HRTF, non-individualized or altered, in a short time period through various training programs going from simple sound exposure to active learning. While all training programs are based on a bimodal coupling (audio-vision or audio-proprioception), they are rarely based on a trimodal one. Our study compares two versions of active trainings: an audio-proprioceptive one and an audio-visuo-proprioceptive one, in order to explicit the role of vision in short-term audio localization training when action and proprioception are already involved. Results from an experimental between-subjects design study, with 27 participants trained on three different program conditions with or without vision, reveal that vision seems to have no or very little influence on rapid audio-proprioceptive trainings and HRTF adaptation. This study could also help a better utilization of 3D-audio for neurological or psychiatric rehabilitation program.

1. INTRODUCTION

Auditory spatial perception and sound source localization rely on several auditory cues (binaural and spectral cues [1]) that allow us to estimate the position of a given sound source. They are contained in the Head Related Transfer Function (HRTF) simulating the transformations caused by the head, the pinna and the torso, for a sound given position. HRTF thus depends on the anatomical features of the listener, and are deeply individualized. Measuring them for each individual user of a VR system is a very long and expensive process. Therefore, it is frequent to use a set of HRTF, measured on another individual or even computationally generated.

Several studies have demonstrated that it is possible to adapt our auditory localization system to new or altered HRTF, for example naturally due to hair cut change and morphologic evolution [2], or with the help of training programs. Three types of training programs, i.e. sound exposition, feedback, and active learning, have been developed. The concept of sound exposition training is based on HRTF modification by artificially altering the outer ear. For example in [3], participants wore earmolds for as long as sixty days. The results showed a significant improvement in sound localization. However, the improvement did not lead to equal performance as those measured with individual HRTF. Feedbacks have also been used to study their effect on the adaptation to new HRTF. [4] used visual feedbacks to provide the correct position of the sound source, after each estimation of its localization. According to [5], feedbacks on the estimation during a localization task improved the performance in a greater measure than a simple exposition to the sound. Lastly, some studies also investigated the effect of active learning. [6] designed a method involving procedural and active learning using visual feedbacks. Once the participants pointed towards the perceived sound, they received a visual feedback at the position of the sound source. They had next to correct their estimation by pointing in the direction of the visual target. Then, to associate the visual and the auditive modalities, both targets were presented at the same time, and had to be pointed again. It is also possible to improve the performance with an active comparison of sounds [7, 8]. Finally, it has been proved that it is possible to shorten the adaptation phase using an active and implicit learning task as suggested by Parseihian and colleagues [9]. Carried out on 3 consecutive days, their training program consisted in a mini sonified version of a hotandcold game where blindfolded participants actively explore the sphere around them to search for invisible targets using a position-tracked ball held in their hand. This game-like task has the advantage to foster the immersion in the audio-virtual environment.

All these training methods are based on multimodal learning. According to [10], these kinds of learning methods are more effective than unimodal learning methods. Furthermore, multimodal learning influences unisensory perception. Indeed, it has been shown that presenting congruent auditory and visual stimuli during the learning stage leads to a greater improvement in visual performance than a visual-only training [11, 12]. Other studies demonstrated that congruent auditory and visual stimuli could also benefit for auditory performance [13]

© Tifanie Bouchara, Tristan-Gaël Bara, Pierre Louis Weiss, Alma Guilbert. Licensed under a Creative Commons Attribution 4.0 International License (CC BY 4.0). **Attribution:** Tifanie Bouchara, Tristan-Gaël Bara, Pierre Louis Weiss, Alma Guilbert. "Influence of vision on short-term sound localization training with non-individualized HRTF", 1st EAA Spatial Audio Signal Processing Symposium, Paris, France, 2019.

doi:10.25836/sasp.2019.04

While all training programs are based on a bimodal coupling (audiovision [14] or audio proprioception [9]), they are rarely based on a trimodal one. It is suprising as many studies have shown that a multisensory learning could be more efficient than a unisensory learning (for a review see [10]). The study [9] has proved that it is possible to rapidly adapt to new HRTF without vision, but it is unclear if vision can reinforce or shorten this adaptation effect. We carried out an experiment presented in the next section to verify our hypothesis that seeing the sound source, during a training program involving implicit learning through an active exploration of the peripersonal sphere, leads to better localization progress than without vision.

2. EXPERIMENTATION

In contrast to earlier studies [9, 15], in which participants were systematically blinded, the aim of our study is to evaluate the influence of vision on sound localization training programs.

2.1 Procedure

2.1.1 Tasks and stimuli

	Day		
	1	2	3
Pre-localization test	L1		
Training task	T1	T2	T3
Post localization test	L2	L3	L4

Table 1. Task presentation across sessions.

As in [9], participants carried out two different tasks: one training task for the adaptation to non-individualized HRTF and one localization task for the assessment of this adaptation. All participants received a training session of 12 minutes during 3 consecutive days and also had to perform 4 sound localization tests: one before the experiment (L1) and one after each training session (L2 to L4).

The adaptation task was similar to the mini-game used in [9] and [15]. Participants had to freely scan the space around them with their hand-held position-tracked Vive controller in order to find an animal sound hidden around them. Target positions were randomly chosen in the frontal hemisphere. The controller-to-target angular distance is sonified through the alternate speed from a white and a pink noise such as the delay between each burst decreased from 3.0 s to 0.05 s with the angular distance (3.0 s meant the target was at the opposite direction). When the target position was reached, the search feedback sound was replaced by a random animal sound. When applicable (groups G_{AP} and G_{AVP}, see experimental design section), the feedback sound and the animal sounds were spatialized through binaural audio at the controller position. Animal sounds were taken from various free sample databases. Participants were asked to find as much as animal sounds they can during the 12 minutes of each adaptation session (T1 to T3, one per day).

Group	Modalities (including proprioception)	HRTF spatialization
Gc	Audio	none
Gap	Audio	non-individual
Gavp	Audio + Vision	non-individual

Table 2. Training task conditions per group.

In the localization tasks (L1 to L4), participants had to report the perceived position of a static spatialized sound sample by pointing with the hand-held controller and validating the direction with the trigger of the controller. As in [9], the stimulus consisted of a train of three 40 ms Gaussian broadband noise bursts (50–20 000 Hz) separated by 30 ms of silence. Each localization test was composed of 2 blocks of 33 trials testing localization performance for 11 azimuths $\{-90°, -72°, -54°, -36°, -18°, 0°, +18°, +36°, +54°, +72°, +90°\}$ × 3 elevations $\{-30°, 0°, +30°\}$. At the end of each new trial, participants first had to point a target presented visually (green object) at a position of $0°$ az., $0°$ el. so participants were always oriented towards the frontal direction at the beginning of a trial. In each block, trials were randomly presented. The mean duration of this task was 3 min per block. Participants were allowed to take a break of 3 min between blocks and tasks. The experiment lasted almost one hour the first day, then 30 minutes on days 2 and 3, for each participant.

2.1.2 Experimental design

We used a between-subjects design where participants were randomly assigned to 3 experimental groups whose conditions are summarized in Table 2.
G_C was a control group, i.e. participants of that group received a training session but it was impossible to gain any HRTF adaptation during that phase because sounds were displayed in mono without any binaural spatialization effect. Participants were still active but any gain in performance from session L1 to L4 could only be attributed to procedural learning effect (due to task repetition) and not to localization or binaural learning. In this condition, searching for hidden target was still possible as it relies on sonification processes and not in sound localization.
G_{AP} received an audio-proprioceptive training as exposed in [9, 15]. The animal sounds and the feedback sound indicating the angular distance to the target were spatialized with binaural. No visual information was provided.
G_{AVP} received the same task as in G_{AP} but, a visual representation of a sphere was also displayed at the hand position during all training sessions (audio-visuo-proprioceptive situation).

2.2 Hypotheses

We hypothesized that: H1) training programs would lead to an auditory adaptation when using implicit active learning and HRTF presentation ; H2) combining all modalities in the audio-visuo-proprioceptive program would optimize

doi:10.25836/sasp.2019.04

the adaptation, inducing better performance and a longer remaining effect than an audio-proprioceptive program.

2.3 Participants

Twenty-seven volunteers (age: 18-43 years (mean 23.5, St. 5.7); 24 females; 25 right-handed), essentially students from undergraduate programs of Paris Descartes University, participated in this between-subject designed study. They were randomly assigned to the experimental groups, finally composed as such: G_C (6 participants, 6 women, all right-handed, mean age 23 (ST 4.8)), G_{AP} (11 participants, 2 men, all right-handed, mean age 23.7 (ST 4.9)), and G_{AVP} (10 participants, 2 men, 8 right-handed, mean age 23.7 (ST 7.3)).

Participants were tested individually in the same isolated listening room, seated in a swivel chair. All participants reported normal or corrected to normal vision and normal hearing. All participants had an audiometric test before the experiment, verifying normal audiometric thresholds (less than 20 dB HL) at octave frequencies between 250 and 8000 Hz, and no history of hearing difficulties. Subjects were naive to the purpose of the experiment and the sets of spatial positions selected for the experiment. Informed consent was obtained from all participants. This experiment was approved by the Paris Descartes University Ethical Committee (CER) (authorization number: $N°2019-24$).

2.4 Apparatus

The experiment was conducted in a controlled lab environment. The audio-virtual environment was developed under Unity with Steam VR, and was rendered using a HTC Vive as a head- and hand-tracker. 3D audio spatialization is obtained through Steam Audio's non-individualized built-in HRTF. When applicable, 3D visual information is displayed directly on the Vive screen. The computer was composed of two-intel core i7-4790K as CPU, two GeForce GTX 980 as GPU, 16GhZ of RAM and a MSI Z97 Gaming 5 motherboard. Open circum-aural reference headphones (Sennheiser HD 380 Pro) were used without any headphone compensation.

3. RESULTS

3.1 Localization task

3.1.1 Dependant variables and analyses methodology

Target and response azimuths and elevations were logged for each trial during the different localization tests. These dependent variables were converted to the interaural polar coordinate system (lateral and polar angle), initially presented in [16], to analyze the type of error (precision, front/back confusion and up/down confusion) as explained in [9] and [15]. The lateral angle ($-90° \leqslant \alpha \leqslant 90°$) was calculated from the median plane to the represented vector ; the polar angle ($-90° \leqslant \alpha \leqslant 270°$) indicates the rotation around the interaural axis, with $0°$ being front. Then all

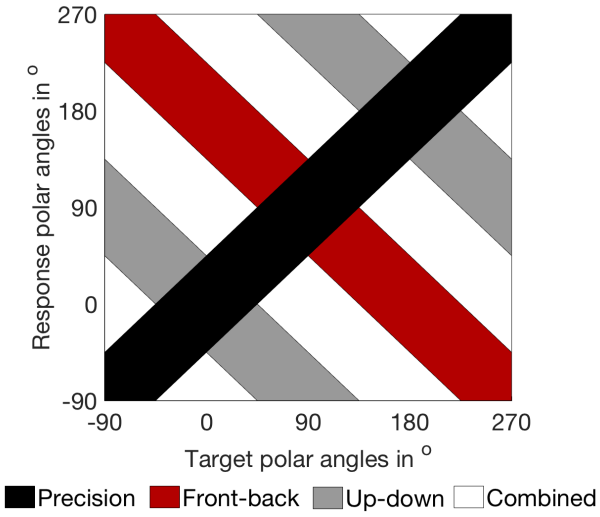


Figure 1. Definition of the four different error type zones according to [9].

error types were determined according to [9] and [15] using the different zones of scatter plot response versus target polar angle as presented in Figure 1.

As the distributions of absolute lateral and polar errors were not normal due to front-back confusions [9], we used the median instead of the mean to analyse the lateral and polar errors. Table 3 reports the average median of localization errors in lateral and polar angles, and the average percentages of front-back confusions for the different groups and localization tests.

A Kruskal-Wallis (non-parametric) ANOVA was carried out on the three dependent variables, lateral and polar errors and percentages of front-back inversion to determine the effect of the group (inter-group comparisons) before the training (localization test L1) and after the training (test L4). A Friedman (non-parametric) ANOVA was carried out on the same three dependent variables to evaluate the adaptation effect, i.e. to determine the effect of the sessions in each group. Wilcoxon tests were carried out to find if there is a significant difference between sessions L1 and L4. Finally, in order to determine which participant showed a learning effect, we carried out intra-subject analyses: Mann-Whitney tests were done on our three dependent variables to compare session L1 to L4. A significance threshold of .05 (one-tailed alpha level) was adopted for all statistical analyses.

3.1.2 Lateral Error

The mean of absolute lateral error medians is shown in Figure 2 for each group over the course of the 4 localization tests (L1 to L4).

Our three groups were not significantly different in session L1 ($H(dl{=}2, N{=}27){=}0.41, p{=}.41$) nor in session L4 ($H(dl{=}2, N{=}27){=}.48, p{=}.39$) for lateral errors.

A significant improvement across sessions was found for G_{AP} ($F(dl{=}2, N{=}11){=}10.2, p{=}.0085$). Participants were better in session L4 than in

		Localization test			
		L1	L2	L3	L4
G_C	Lateral Error Medians	18.3 (3.8)	15.8 (3.2)	16.0 (3.8)	17.1 (5.7)
	Polar Error Medians	88.6 (65.7)	82.6 (57.4)	86.5 (56.1)	76.1 (54.1)
	Percent of Front-Back Error	28.3 (20.9)	21.4 (18.7)	23.5 21.8)	21.0 (23.2)
G_{AP}	Lateral Error Medians	19.9 (5.0)	19.5 (4.1)	17.6 (5.6)	16.0 (3.8)
	Polar Error Medians	58.0 (27.9)	66.7 (41.6)	64.4 (39.6)	60.2 (45.4)
	Percent of Front-Back Error	21.8 (14.5)	19.0 (16.4)	20.2 (19.5)	17.8 (19.9)
G_{AVP}	Lateral Error Medians	18.8 (7.4)	17.9 (7.4)	17.9 (7.4)	17.3 (5.2)
	Polar Error Medians	107.7 (60.6)	103.3 (64.0)	95.7 (59.5)	102.5 (65.3)
	Percent of Front-Back Error	29.7 (24.4)	29.1 (24.8)	24.6 (22.6)	26.1 (23.0)

Table 3. Mean of lateral error medians (in °), mean of polar error medians (in °), and percentages of front-back errors across session and groups. Numbers in brackets indicate standard deviations.

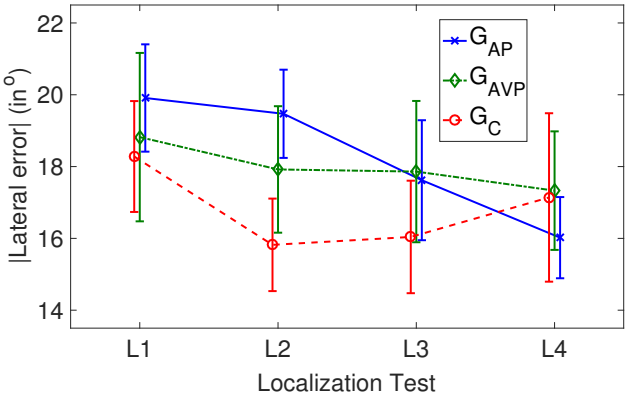


Figure 2. Mean absolute lateral error medians during the four localization tests for each group (control without HRTF presentation, audio-proprioceptive and audio-visuo-proprioceptive). Vertical bars indicate standard error of the mean.

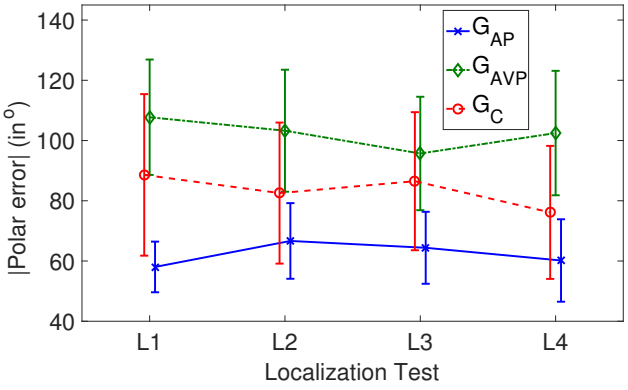


Figure 3. Mean absolute polar error medians during the four localization tests for each group (control without HRTF presentation, audio-proprioceptive and audio-visuo-proprioceptive). Vertical bars indicate standard error of the mean.

session L1 (Z=2.67, p=.0038). No significant improvement across the sessions was found for G_{AVP} ($F(dl$=2, N=10)=0.72, p=.43) and G_C ($F(dl$=2, N=6)=2.69, p=.22). Again, no improvement between session 1 and session 4 was found for G_{AVP} (Z=0.76, p=.22) and G_C (Z=0.52, p=.30).

Intra-subject analyses revealed that 5/11 participants from G_{AP}, 5/10 from G_{AVP}, and 3/6 from G_C were or tended to be better for session L4 than for session L1.

3.1.3 Polar Error

For both sessions L1 and L4, our three groups were not significantly different in polar errors (L1: $H(dl$=2, N=27)=2.74, p=.13; L4: $H(dl$=2, N=27)=2.30, p=.16). However there was a statistical tendency to a difference between G_{AP} and G_{AVP} in both session L1 before learning (U=32, Z= − 1.58, p=.057) and session L4 after learning (U=35, Z= − 1.37, p=.087).

No significant effect of the session was found neither for G_{AP} ($F(dl$=2, N=11)=0.054, p=.5) nor for G_C ($F(dl$=2, N=6)=3.2, p=.18). Only a statistical tendency for an effect of session was found for the group G_{AVP} ($F(dl$=2, N=10)=5.88, p=.059). However this was not due to learning as the difference between L1 and L4 for that group G_{AVP} was not significant (Z=0.005, p=.48).

Intra-subject analyses of absolute polar errors revealed that 5/11 participants from G_{AP}, 1/10 from G_{AVP} and 2/6 from G_C were or tended to be better for session L4 than for session L1.

3.1.4 Front-back confusion

Our three groups were not significantly different in percentage of front-back inversion in session L1 ($H(dl$=2, N=27)=0.35, p=.42) nor for session L4 ($H(dl$=2, N=27)=0.42, p=.40).

No significant improvement across sessions was found for none of the three groups G_{AP} ($F(dl$=2, N=11)=1.78, p=.31), G_{AVP} ($F(dl$=2, N=10)=2.6, p=.23) and G_C ($F(dl$=2, N=6)=3.86, p=.28). Only a statistical tendency between session L1 to L4 was found for the participants of group G_{AP} (Z=1.42, p=.077) with a trend to less front-back confusion errors at the end of the training.

doi:10.25836/sasp.2019.04

3.2 Adaptation task

No measure of localization abilities was done in this task. However, the number of sound sources found during this task can provide indications on the task difficulty. First, a comparison between the three trainings (T1 to T3) was carried out to underline the progress across trainings thanks to a Friedmann (non-parametric) ANOVA and Wilcoxon tests. Then, a comparison between our three groups was made thanks to a Kruskal-Wallis (non-parametric) ANOVA in order to determine if the use of HRTF or vision could make the task easier. Table 4 reports the average of completed trials per group, i.e. the number of animals found, during each session of 12 min.

The number of sound sources found was different across the sessions ($F(dl{=}2, N{=}27){=}29.7, p{<}.0001$). Participants performed better during T2 than during T1 ($Z{=}3.34, p{=}.0004$) and during T3 than during T2 ($Z{=}3.1, p{=}.001$). Performance improved across the sessions.

Our three groups did not show significant differences during any of the training ($T1{:}H(dl{=}2, N{=}27){=}0.21, p{=}.45$; $T2{:}H(dl{=}2, N{=}27){=}0.93, p{=}.31$; $T3{:}H(dl{=}2, N{=}27){=}0.075, p{=}.48$). The use of HRTF or vision did not seem to make the task easier.

	Training session		
	T1	T2	T3
G_C	13.2 (1.1)	15.0 (2.7)	20.2 (1.4)
G_{AP}	14.3 (8.1)	17.7 (7.2)	19.8 (7.0)
G_{AVP}	12.4 (6.1)	18.8 (6.9)	20.5 (5.4)

Table 4. Mean number of animals found per session (standard deviation) across session and groups.

4. DISCUSSION

The only significant improvement of localization performance across the sessions was observed for the audio-proprioceptive group G_{AP}: a significant learning effect was underlined on lateral errors between before and after the program. That means that our training program based on an active and implicit audio-proprioceptive learning task is efficient for non-individualized HRTF adaptation, as shown by [9]. No significant improvement was observed for the audio-visuo-proprioceptive group G_{AVP} and for the control group G_C. Vision did not lead to better HRTF localization performance in our study, suggesting that the program without vision could be better than the one with vision. This does not support our hypothesis of a better localization progress with vision than without.

Nevertheless, this result could be explained by some particularities of multisensory integration. As seen previously, multisensory learning could be more efficient than a unisensory learning on the perception of a particular modality [11–13]. Spatial and temporal coincidences facilitate multisensory integration [17]. Hypotheses state that discrepancies are resolved in favor of the more appropriate modality [14]. Although hearing is predominant for temporal perception, vision is more precise for spatial judgments. However, in our experiment, although the visual information is spatially and temporally congruent with the auditory one, vision did not seem to improve the learning of the spatial positions of the new HRTF. This absence of effect could be explained by the fact that multisensory integration depends on some other conditions. One of the main principles underlying multisensory integration is the inverse effectiveness rule [17]. According to this principle, the efficiency of the integration depends on the nature of stimuli: the more unimodal stimuli are ambiguous and weak, the more another modality is used, even if the information from this modality is not apparently related to the task [18]. In our adaptation task, vision was informative to localize the hand position in space. However, the visual information did not inform of the distance from the target, contrary to the auditory information. The auditory information may have appeared sufficient to come up with a robust estimate and realize the task and, thus, information from several modalities was not combined, such as suggested by the probabilistic model of [19], and visual information could even be distracting for the participants. One way to provoke more multisensory integration would be to give less information to the auditory modality and more to the visual one. We could imagine, in a future study, to give the information of distance from the target also to the visual modality in order to show if we could improve the visual contribution and, thus, improve the learning of new HRTF.

However, in our study, no difference between the three groups of participants can be observed at the end of the third and last training session. This could be due to a sample of participants being too small to show statistical significance. This could also be due to an insufficient number of learning sessions. Indeed, the audio-proprioceptive group GAP especially improved after the third training session. Therefore, we could hypothesize that the adaptation will be greater after a fourth session or more. This is coherent with the results of recent studies [15] who have shown a continuing improvement over a program of 10 weeks, one session per week. Another explanation would be that groups may be heterogeneous in terms of matching between the individual participant HRTF and the non-individualized generic HRTF. Indeed, the adaptation slope depends on the compatibility between the HRTF of the listener and the HRTF to be learnt [9] and, in each of our groups, some participants were able to improve their localization performance whereas others were not. Moreover, the audio-proprioceptive group G_{AP} and the audio-visuo-proprioceptive group G_{AVP} tended to be different in polar errors before learning. This suggests that our two groups may have had differences in terms of matching with the non-individualized generic HRTF before the learning. Another experiment with more participants and a preselection step to select compatible HRTF for all of these participants is finally required to

really assess if vision can or not improve adaptation to HRTF.

Another point that needs to be highlighted is that three participants from the control group improved across the sessions on the lateral and/or on the polar errors. Yet, this group was not exposed to any HRTF during the training. So, the statistical improvements can only be explained to a familiarization with the material and the task. This definitely is an argument to encourage further HRTF training studies to always compare with a control group performing the task in mono instead of comparing with a group receiving no training at all.

Finally, our study and future ones could improve the use of virtual reality and 3D audio as a rehabilitation strategy for specific listeners suffering from neurological disorders, such as spatial cognition disorders. For example, our program could be adapted for unilateral spatial neglect, which is a common neurological disorder in which patients have difficulties to pay attention to the contralesional side of space in vision, but also in other sensory modalities such as hearing [20].

5. REFERENCES

[1] J. Blauert, *Spatial Hearing, The Psychophysics of Human Sound Localization*. MIT Press, Oct. 1996.

[2] J. P. Rauschecker, "Auditory cortical plasticity: a comparison with other sensory systems.," *Trends neurosci.*, vol. 22, pp. 74–80, feb 1999.

[3] S. Carlile, K. Balachandar, and H. Kelly, "Accommodating to new ears: The effects of sensory and sensory-motor feedback," *J. Acoust. Soc. Am.*, vol. 135, pp. 2002–2011, apr 2014.

[4] B. G. Shinn-Cunningham, N. I. Durlach, and R. M. Held, "Adapting to supernormal auditory localization cues. I. Bias and resolution," *J. Acoust. Soc. Am.*, vol. 103, pp. 3656–3666, jun 1998.

[5] K. Strelnikov, M. Rosito, and P. Barone, "Effect of Audiovisual Training on Monaural Spatial Hearing in Horizontal Plane," *PLOS ONE*, vol. 6, pp. 1–9, 03 2011.

[6] P. Majdak, M. J. Goupell, and B. Laback, "3-d localization of virtual sound sources: Effects of visual environment, pointing method, and training," *Atten. Percept. Psycho.*, vol. 72, pp. 454–469, Feb 2010.

[7] C. Mendonça, G. Campos, P. Dias, J. Vieira, J. P. Ferreira, and J. A. Santos, "On the improvement of localization accuracy with non-individualized HRTF-based sounds," *J. Audio Eng. Soc.*, vol. 60, pp. 821–830, Oct. 2012.

[8] C. Mendonça, G. Campos, P. Dias, and J. A. Santos, "Learning auditory space: Generalization and long-term effects," *PLOS ONE*, vol. 8, 10 2013.

[9] G. Parseihian and B. F. G. Katz, "Rapid head-related transfer function adaptation using a virtual auditory environment," *J. Acoust. Soc. Am.*, vol. 131, no. 4, pp. 2948–2957, 2012.

[10] L. Shams and A. R. Seitz, "Benefits of multisensory learning," *Trends Cogn. Sci.*, vol. 12, no. 11, pp. 411–417, 2008.

[11] A. R. Seitz, R. Kim, and L. Shams, "Sound Facilitates Visual Learning," *Curr. Biol.*, vol. 16, pp. 1422–1427, jul 2006.

[12] R. S. Kim, A. R. Seitz, and L. Shams, "Benefits of stimulus congruency for multisensory facilitation of visual learning," *PLOS ONE*, vol. 3, pp. 1–5, 01 2008.

[13] K. von Kriegstein and A.-L. Giraud, "Implicit multisensory associations influence voice recognition," *PLOS Biology*, vol. 4, pp. 1–12, 09 2006.

[14] C. Mendonça, "A review on auditory space adaptations to altered head-related cues," *Front. Neurosci.*, vol. 8, p. 219, 2014.

[15] P. Stitt, L. Picinali, and B. F. Katz, "Auditory Accommodation to Poorly Matched Non-Individual Spectral Localization Cues Through Active Learning," *Sci. Rep.*, vol. 9, no. 1, 2019.

[16] M. Morimoto and H. Aokata, "Localization cues of sound sources in the upper hemisphere.," *J. Acoust. Soc. Japan (E)*, vol. 5, no. 3, pp. 165–173, 1984.

[17] B. E. Stein and M. A. Meredith, *The merging of the senses*. Cambridge , MA, The MIT Press, Jan. 1993.

[18] E. B. Stein, N. London, K. L. Wilkinson, and P. Donald, "Enhancement of perceived visual intensity by auditory stimuli: A psychophysical analysis," *J. Cogn. Neurosci.*, vol. 8, pp. 497–506, 11 1996.

[19] M. O. Ernst and H. H. Bülthoff, "Merging the senses into a robust percept," *Trends Cogn. Sci.*, vol. 8, no. 4, pp. 162 – 169, 2004.

[20] A. Guilbert, S. Clément, L. Senouci, S. Pontzeele, Y. Martin, and C. Moroni, "Auditory lateralisation deficits in neglect patients," *Neuropsychologia*, vol. 85, pp. 177 – 183, 2016.

 doi:10.25836/sasp.2019.04

THE INFLUENCE OF DIFFERENT BRIR MODIFICATION TECHNIQUES ON EXTERNALIZATION AND SOUND QUALITY

Peter Maximilian Giller **Florian Wendt** **Robert Höldrich**
Institute of Electronic Music and Acoustics, University of Music and Performing Arts Graz
giller@student.tugraz.at, {wendt|hoeldrich}@iem.at

ABSTRACT

In the context of binaural audio, externalization refers to the sensation of virtual sound sources being located outside of the listener's head. Binaural reproduction using anechoic head-related impulse responses is known to suffer from poor externalization. The degree of externalization can be increased by reverberation, as contained in binaural room impulse responses. However, the presence of reverberation is not always desired since the original sound of a recording should usually be preserved. This study concerns the dilemma of creating well-externalized dry-sounding signals. We investigated the manipulation of either the impulse response length, the reverberation time, or the direct-to-reverberant energy ratio regarding externalization and attributes of sound quality. As expected, each condition is a compromise between externalization and sound quality. While externalization increases with increasing amount of reverberation for all methods in a similar way, our findings show that the differences between them lie in sound color and perceived naturalness.

1. INTRODUCTION

Binaural synthesis aims to produce well externalized sound images, i.e., auditory images are perceived to be located outside the listener's head, 'compact and correctly located in space' [1]. The perceptual and technical aspects of the phenomenon were the subject of earlier research [1–6]. Studies showed that the presence of reverberation can increase the degree of externalization [2, 3], but also introduce sound colorations [4]. While, for binaural rendering, it is usually desired to preserve the room impression and sound color of the original recording, this may lead to a conflict between the synthesized and the listening room known as the room divergence effect [5].

We distinguish head-related impulse responses (HRIRs), which capture the influence of reflections at the pinnae, head, and torso from the direct sound impinging at the ear canals from a certain direction, and binaural room impulse responses (BRIRs). The latter consist of the direct part, which is identical to the HRIR, followed by early reflections and diffuse reverberation from all directions, convolved with the HRIRs for the respective directions.

The goal is to add as little reverberation as necessary to HRIRs in order to increase externalization, or to reduce the reverberation of BRIRs as much as possible in order to reduce differences in sound. The influence of BRIR truncation, the simplest method to reduce reverberation, on externalization was investigated in several studies [2–4]. It was found that a minimum BRIR length of 80-100 ms is sufficient to yield externalized sound images [2, 3]. Since there is no equivalent physical phenomenon leading to truncated BRIRs, this method yields a rather artificial sound. In contrast, natural de-reverberation can either be achieved by increasing the absorption area in a room, yielding a shorter reverberation time, or by reducing the source distance, leading to a higher direct-to-reverberant energy ratio (DRR).

In this study, we investigate the influence of modifications of the reverberant part systematically in order to establish a connection between the achieved externalization and sound quality. Three basic approaches are studied, which are: (*i*) truncation of the impulse response length (temporal modification), (*ii*) manipulation of the DRR by weighting the reverberant part with a constant factor (modification of level), and (*iii*) alteration of the reverberation time by weighting the reverberant part with an exponential decay function (temporal modification of level). Section 2 describes the above mentioned modification methods. We have compared the three approaches in a listening experiment introduced in Section 3. The participants compared the degree of externalization as well as sound quality by rating naturalness and similarity to an anechoic signal. Section 4 discusses the results of the experiment.

2. MANIPULATION OF IMPULSE RESPONSES

Fig. 1 gives an illustrative example of the studied modification techniques. The dashed black line represents the envelope of a BRIR $h(t)$ on a logarithmic scale. The vertical dashed line marks the boundary between the direct part $h_{\mathrm{dir}}(t)$, which we define as the direct sound followed by all reflections from pinnae, head, and torso (the HRIR), and the reverberant part $h_{\mathrm{rev}}(t)$, containing early and diffuse reflections from the surroundings, where $h = h_{\mathrm{dir}} + h_{\mathrm{rev}}$. Each of the modifications can be understood as a time-variant weighting of each impulse response $h(t)$, as illustrated in

© Peter Maximilian Giller, Florian Wendt, Robert Höldrich. Licensed under a Creative Commons Attribution 4.0 International License (CC BY 4.0). **Attribution:** Peter Maximilian Giller, Florian Wendt, Robert Höldrich. "The influence of different BRIR modification techniques on externalization and sound quality", 1st EAA Spatial Audio Signal Processing Symposium, Paris, France, 2019.

doi:10.25836/sasp.2019.20

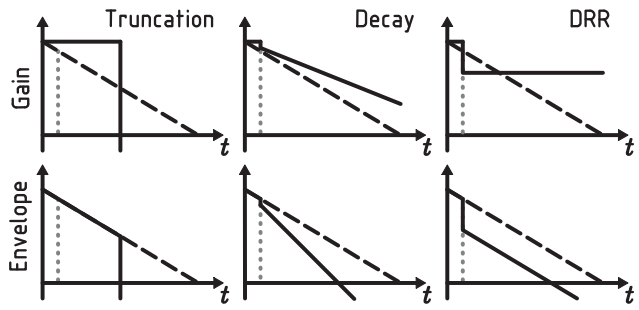


Figure 1. Illustration of the modification methods. The black and gray dashed lines show the envelope of the BRIR and the boundary between direct and reverberant part. The upper row shows the equivalent time-dependent gain, whereas the lower row shows the resulting envelope.

the upper row. The lower row of Fig. 1 illustrates the resulting envelopes.

As shown in the left column, the truncation method applies a window of length L with unit gain to the BRIR, with a short fade in the transition region where needed.

In contrast, by multiplication with an exponential decay curve the reverberation time is modified, as shown in the center column. The decay curve begins at the position of the direct sound, but, in order to leave the HRIR unaffected, is only applied to the reverberant part of the BRIR,

$$\tilde{h}_{\mathrm{rev}}(t) = h_{\mathrm{rev}}(t) \cdot 10^{-\frac{60}{20}(\tilde{T}_{30}^{-1} - T_{30}^{-1})t}, \qquad (1)$$

where T_{30} is the reverberation time estimated from the original BRIR, and $\tilde{T}_{30}$ is the target reverberation time. To realize the weighting, the BRIR is split into its direct and reverberant parts, using a short overlapping window in the transition region, which are then individually weighted.

The right column of Fig. 1 exemplifies the manipulation of the DRR, which is defined as

$$DRR = 10\lg\left(\frac{\int h_{\mathrm{dir}}^2(t)\mathrm{dt}}{\int h_{\mathrm{rev}}^2(t)\mathrm{dt}}\right). \qquad (2)$$

It can be manipulated if the reverberant part of the BRIR, split analogously to the decay method, is multiplied with a constant factor.

3. LISTENING EXPERIMENT

A listening experiment was conducted to evaluate the impact of the three modifications truncation, decay, and DRR on externalization and sound quality. Externalization is known to be particularly fragile for the frontal direction which is why we investigated speech sources simulated using non-individual BRIRs direct in front of the listeners at $0°$, played back over headphones. We did not carry out individual measurements because the experiment concerns rather the relative differences to the original, unmodified BRIR than the absolute differences to a physical sound source.

The experiment consisted of three consecutive parts: The first part (part *N*) was an exploratory evaluation of the perceived naturalness of a subset of the created conditions. In the second part, participants had to rate the externalization (part *E*), and in the last part sound quality was evaluated by rating the similarity to an anechoic signal regarding different attributes of sound (part *S*).

Method		**Truncation**	**Decay**	**DRR**
Parameter		L (s)	T_{30} (s)	DRR (dB)
Condition i	0 (BRIR)	1.0	0.70	2.3
	1	0.350	0.60	5.3
	2	0.193	0.51	8.3
	3	0.106	0.42	11.3
	4	0.059	0.33	14.3
	5	0.032	0.23	17.3
	6	0.018	0.14	20.3
	7	0.009	0.05	23.3
	8 (HRIR)	0.003	-	∞

Table 1. Conditions created within each method of modification.

3.1 Conditions

Virtual sounds sources were created using a BRIR measurement of a Neumann KU 100 dummy head. This measurement constitutes the starting point for the creation of all other conditions and was obtained in the lecture room of our institute (dimensions $3\,\mathrm{m} \times 7\,\mathrm{m} \times 8.3\,\mathrm{m}$, reverberation time $T_{30} = 0.7\,\mathrm{s}$). The sound source, a Neumann KH 120 loudspeaker, was located at a distance of $2.5\,\mathrm{m}$ to the receiver, at an angle of $0°$. Both source and receiver were positioned at a height of $1.25\,\mathrm{m}$. To provide additional reflections from breast and shoulders, possibly supporting externalization, the dummy head was equipped with the torso of a Brüel & Kjær HATS. Different conditions were created by manipulating the length, reverberation time, or DRR of the measured BRIR. An anechoic $8\,\mathrm{s}$ long sequence of male speech was convolved with the resulting BRIRs corresponding to each of the created conditions.

Tab. 1 lists the parameter values of each condition. The condition with index $i = 0$ corresponds to the original, unmodified BRIR, whereas the HRIR, i.e., the BRIR truncated immediately before the arrival of the first reflection from the room, has the index $i = 8$. The parameter levels for each modification were selected heuristically based on the experience from preceding informal experiments of the authors to achieve, with regard to reverberation, a near-uniform sampling of the parameter ranges from unmodified BRIR to the HRIR, i.e., from congruence to divergence between the synthesized and the real room.

3.2 Playback and Equalization

All conditions were played back via headphones, except for the loudspeaker reference condition in the externalization part *E*. To facilitate comparative rating, participants wore open headphones throughout the whole experiment. Unfortunately, the headphone alters the sound from the loudspeaker reference somewhat as the sound has to propagate through the ear cups. The AKG K702 headphone was thus modified in order to reduce the damping of frontal sound as

doi:10.25836/sasp.2019.20

much as possible by replacing the ear cushions by self-made ones with cutouts at the front and back. Remaining differences in sound color between headphone and loudspeaker were equalized using a minimum-phase filter. Far-field responses from the loudspeaker to the dummy head, both with and without the modified headphone, were measured in an anechoic chamber. The magnitude spectra were smoothed within critical bands and the filter for the headphone signal was obtained by dividing the without-headphones magnitude frequency response by the with-headphones response.

Obviously, the modification of ear cushions distorts the frequency response of the headphone itself. To linearize the magnitude transfer function from each headphone driver to the corresponding ear canal, the headphones signals were convolved with an additional minimum-phase inverse filter of the magnitude spectrum smoothed within critical bands. This second equalization also removes the undesired contribution of the pinnae and ear canals (of the dummy head) to the headphone transfer function.

3.3 Experimental Design

Each part of the listening experiment was carried out in a MUSHRA-like procedure where participants had to rate a number of conditions of the three different modification techniques. Participants were asked to rate every condition in the presented set of stimuli with continuous sliders on a graphical user interface. They were allowed to repeat each condition at will, and audio files were played back in loop.

The configuration of the test is summarized in Tab. 2. Parts *E* and *S* each consisted of two stages. The first stage (*E*.I/*S*.I) we refer to as the *indirect* comparison. It consists of three sets presented in random order, each corresponding to one modification technique, i.e., one column of Tab. 1. While this setup enables us to draw a comparison between conditions of each modification, a cross-comparison between modifications can solely be achieved indirectly via comparison to the common reference, hidden reference, and anchor. Therefore, in the second stage (*E*.II/*S*.II), a *direct* cross-comparison of modifications was carried out – due to the large number of conditions only with a subset of conditions $i = \{2, 4, 6\}$ from each modification in the same set. These conditions were selected by informal listening to yield preferably similar ratings of externalization or similarity for all modifications in each of the three corresponding levels.

The ratings from the direct comparison x_i^{II} for conditions $i = \{0, 2, 4, 6, 8\}$ were then used to obtain complete corrected curves x_i for each listener by linear scaling and shifting of the ratings of the indirect comparison x_i^{I} for conditions $i = \{1, 3, 5, 7\}$. With $i = 1...8$, this yields a complete set of ratings

$$x_i = \begin{cases} x_i^{\mathrm{II}} & \text{even } i, \\ x_{i-1}^{\mathrm{II}} + \frac{x_{i+1}^{\mathrm{II}} - x_{i-1}^{\mathrm{II}}}{x_{i+1}^{\mathrm{I}} - x_{i-1}^{\mathrm{I}}} \left(x_i^{\mathrm{I}} - x_{i-1}^{\mathrm{I}}\right) & \text{odd } i \end{cases} \quad (3)$$

per listener for each method, allowing for a cross-comparison between the methods.

Part		Method (abbreviation)	Reference	Hidden Reference	Conditions	Anchor
N		all	-	0	2,4,6	8
E	I	Truncation (*trc*)	LS	0	1-7	8
		Decay (*dec*)	LS	0	1-7	8
		DRR (*drr*)	LS	0	1-7	8
	II	all	LS	0	2,4,6	8
S	I	Truncation (*trc*)	8	8	1-7	0
		Decay (*dec*)	8	8	1-7	0
		DRR (*drr*)	8	8	1-7	0
	II	all	8	8	2,4,6	0
	IIa-b	all	8	8	2,4,6	0

Table 2. Experimental setup. Numbers correspond to the condition indices in Tab. 1, and *LS* refers to the loudspeaker.

Part		Task: 'Rate the...
N		...naturalness!'
E	I-II	...externalization, compared to the reference!'
S	I-II	...similarity to the reference *in general*!'
	IIa	...similarity regarding *sound color*!'
	IIb	...similarity reg. the *amount of reverberation*!'

Table 3. Task definitions of each part of the experiment.

The first two parts, *N* and *E*, were conducted in the original room at the position of the measurement. The task definitions of each part are listed in Tab. 3. Part *N* was conducted without a reference, and, thus, the loudspeaker did not play, on which the participants were informed. The participants were asked to rate the naturalness in general, not necessarily bound to the particular room and distance to the loudspeaker. The rating had to be entered on a scale from 0 ('very unnatural') to 100 ('entirely natural'). The stimuli were identical to *E*.II. In part *E*, the loudspeaker was used as a reference to remind the participants of the impression of full externalization of a distant and compact physical sound source. The signal convolved with the BRIR (cond. 0) was used as a hidden reference. The participants were asked to rate the externalization compared to the loudspeaker on a scale from 0 ('inside head') over 33 ('close to the head') to 100 ('at the position of the loudspeaker'). They were instructed not to move their head during playback.

In part *S*.I-II, participants were asked to rate the general similarity to a reference on a scale from 0 ('very different') to 100 ('identical'). In addition, they were asked to rate the similarity regarding sound color or reverberation (*S*.II.a-b) for the reduced set. The reference and hidden reference were the anechoic speech signal convolved with the HRIR (cond. 8). The conditions used were the same as in the first two parts. In order to avoid the influence of spatial attributes on the rating, all stimuli including the reference were presented monaurally by playing back the left-ear signal for both ears. Furthermore, it was conducted in an anechoic chamber to decouple the rating from the measurement room. The stimuli were presented over the modified headphones.

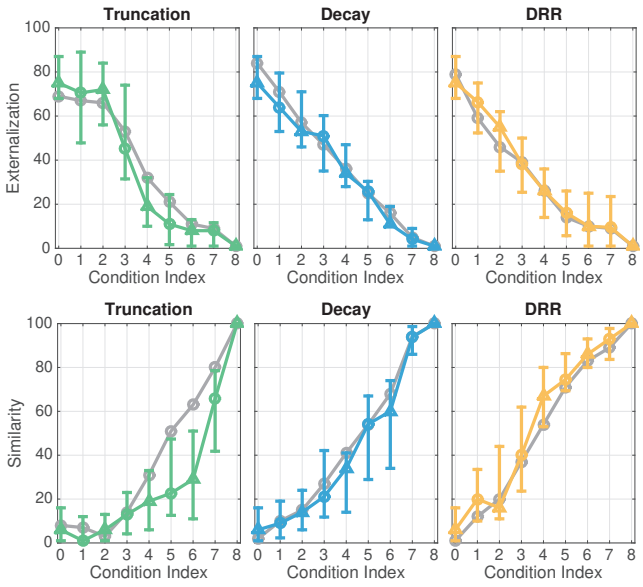


Figure 2. Median original *(gray)*, as well as the scaled ratings *(colored)* with 95 % confidence intervals. The condition indices refer to Tab. 1. Externalization is shown in the upper, and similarity in the lower row. Triangular shapes mark conditions of the direct comparison.

4. RESULTS

Twenty-one experienced listeners participated in the experiment. A Friedman test showed that the effect of the BRIR modification is significant with $p < 0.05$ within every part of the experiment. Fig. 2 shows the median ratings of experiments *E*.I-II and *S*.I-II for each of the modifications truncation, decay, and DRR. For brevity, we will refer to the methods simply as *trc*, *dec*, and *drr* in the following, with, e.g., *dec* i denoting the i-th condition of decay modifications. The externalization ratings are shown in the upper and the similarity ratings in the lower row. The results of the indirect comparisons, *E*.I and *S*.I, are plotted as gray lines in the background. Colored curves represent the ratings corrected based on the direct comparison, *E*.II and *S*.II. The conditions tested in the direct comparison are marked with triangular shapes.

As expected, the overall relation between modification depth and externalization as well as similarity is monotonic, where increasing modification depth leads to a decrease in externalization and an increase in similarity to the HRIR.

4.1 Representation on a Common Axis

The ratings of the different modifications are not directly comparable due to the different nature of the respective varied parameter. In order to relate them to a common physical measure, we computed the temporal centroid of the impulse responses of all conditions. The impulse responses were priorly weighted with the average spectrum of the anechoic speech signal order to limit the evaluation to the frequency content presented to the participants. Fig. 3 shows the scaled externalization and similarity ratings against the temporal centroid on the horizontal axis. Note that the exter-

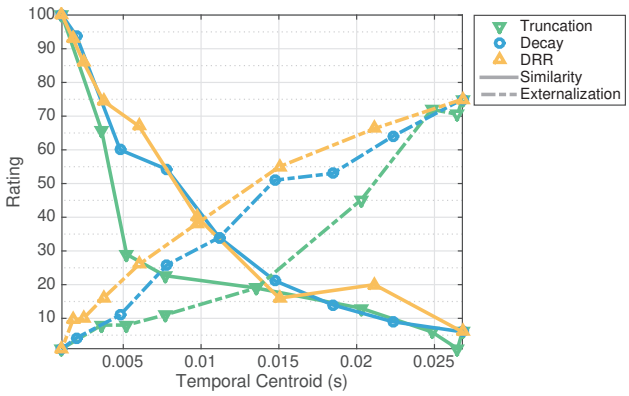


Figure 3. Median scaled ratings of externalization and similarity over the temporal centroid of the weighted BRIRs of each condition.

nalization curves almost describe a straight line for *dec* and *drr*, indicating a good prediction for the methods that have a physical equivalent. Note also the downward deviation of the curve for *trc*, showing that similarly externalized conditions are associated with more late energy in the BRIR. The curves for similarity do not differ much between *dec* and *drr*, but, again, the curve for *trc* deviates towards lower ratings. The temporal centroid is primarily unrelated to fluctuations of binaural cues or timbral properties, which are known to have an effect on externalization [2, 6]. Thus, we do not claim it to be the best approach for a fair comparison. However, it appears to be a suitable measure to connect the ratings to the physical properties of the signal content and reveal differences between the methods therein.

In the following, we will only consider the ratings obtained in the direct comparison, since ratings for naturalness and the sub-attributes of similarity were recorded with the reduced stimulus set for simplicity reasons. The ratings of parts *N* and *S*.II-IIb are shown in Fig. 4. In order to compare each of these attributes to the corresponding externalization rating, *E*.II is shown in the background as a gray line. We performed paired comparisons between the conditions of each part using the Wilcoxon signed-rank test.

4.2 Externalization

As expected, the full BRIR (cond. 0) is externalized best, whereas the HRIR (cond. 8) is not externalized. All other conditions are rated significantly lower than the BRIR. The relation between externalization and the degree of modification, i.e., the amount of reverberation that is being removed from the BRIR, is monotonic within each method. Paired comparison showed that the differences of neighboring levels within one modification, as well as the differences of any of the levels to the full BRIR or the HRIR, are significant.

The BRIR yielded median ratings of only around 80 %. We attribute the lower ratings mainly to individual differences between the listeners' and the employed generic HRIR, as well as remaining timbral differences since equalization was carried for the dummy head. With the loudspeaker available for comparison, the setup was particularly sensitive.

doi:10.25836/sasp.2019.20

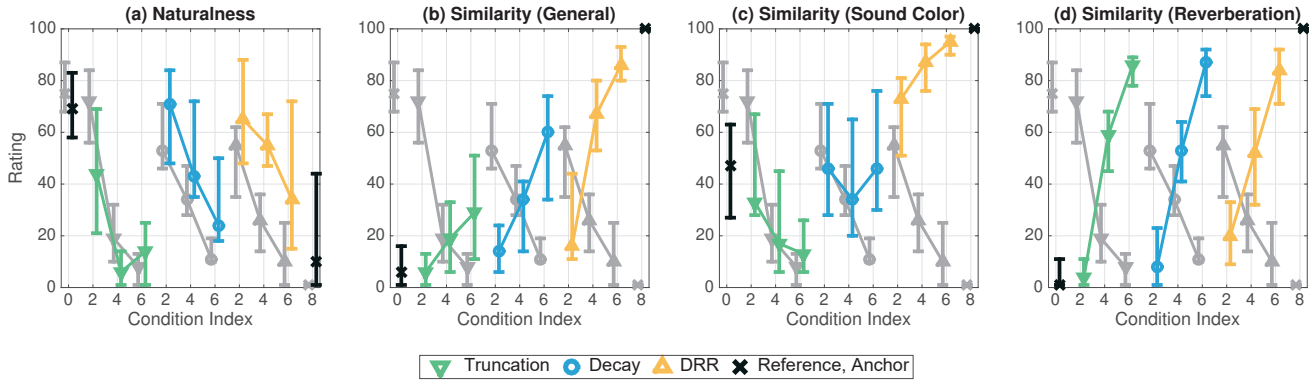


Figure 4. Median and 95 % confidence intervals of the ratings of the direct comparison (*E*.II and *S*.II-IIb). Gray lines in the background represent externalization, and colored lines in the foreground naturalness and similarity to the HRIR. The condition indices refer to Tab. 1.

4.3 Naturalness

The rating of naturalness in Fig. 4(a) similarly tends to decrease monotonically towards less reverberant conditions. Monotony appears to be interrupted between *trc* 4 and 6, but the difference is not significant. Truncation yields lower ratings than the other methods: The maximum rating is significantly lower than the maximum ratings of the other methods. All three conditions of *trc* do not significantly differ from the HRIR, whereas all conditions of the *drr* method and *dec* 2 and 4 do. Furthermore, all truncated conditions are rated significantly lower than the the BRIR, whereas for decay and DRR only cond. 6 is. The rating of the HRIR exhibits a high variation, indicating that anechoic conditions are not necessarily perceived unnatural.

4.4 Similarity

The general similarity (Fig. 4(b)) to speech convolved with the HRIR increases monotonically towards less reverberant conditions within each method. The differences between the levels of each method are significant. All conditions but *trc* 2 differ significantly from the BRIR in their rating. Cond. 2 exhibits the best externalization ratings for each of the modifications, and *trc* 2 is rated significantly better externalized than all conditions but the BRIR. It received a lower similarity rating than the other two methods, too. An increase in externalization with decreasing similarity to the dry reference was expected. However, there appear to be differences between the modifications, as *dec* 4 was rated higher than *trc* 4 in similarity, despite also being rated higher in externalization. The same is true for *drr* 4 and *trc* 6.

The similarity to the HRIR regarding sound color is shown in Fig. 4(c). While the DRR ratings increase monotonically and significantly towards less externalized conditions, the truncation ratings actually decrease with a significant difference between levels *trc* 2, 4, and 6. Moreover, all conditions of *drr* are rated significantly better than all conditions of *trc*. Although no distinct trend is visible for the decay method, *dec* 6 is rated significantly higher than *dec* 4. While none of the ratings of *dec* differs significantly

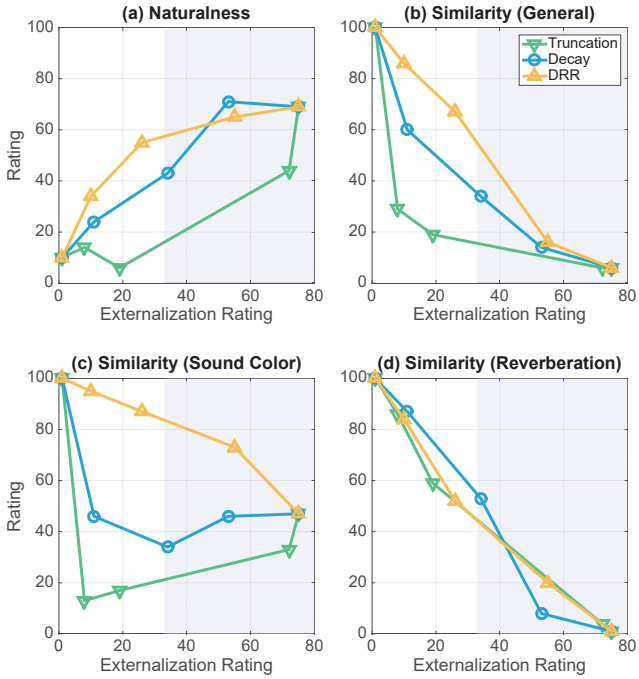


Figure 5. Median of the ratings of the direct comparison, where the naturalness and similarity ratings (parts *N* and *S*.II-IIb) are plotted against externalization (*E*.II). The highlighted area marks conditions considered externalized.

from the BRIR, *trc* 4 and 6 are rated lower than, and all of the DRR modifications are rated higher than the BRIR.

Fig. 4(d) shows the similarity regarding reverberation. At first view, cond. 4 and 6 of each method are perceived similarly reverberant. The ratings within each method increase monotonically and significantly with modification depth (towards less externalized conditions). Yet, in contrast to sound color, *trc* 2 is rated lower than cond. 2 of the other methods. All conditions are rated significantly higher than the BRIR, except for *trc* 2 and *dec* 2. While, again, *drr* 2 is rated higher than *trc* 2, the differences are less pronounced than for sound color. The similarity ratings of the truncation method with regard to sound color decrease

doi:10.25836/sasp.2019.20

while increasing with regard to reverberation, from which it may be inferred that the difference in similarity is due to timbral artifacts. The increase within the DRR method for both attributes may indicate that this method is less likely to produce those artifacts.

4.5 Discussion

We gain a more intuitive view on the data by plotting naturalness and similarity against externalization on the horizontal axis (Fig. 5). The highlighted area marks externalized conditions, i.e., ratings exceeding the 'close to the head' threshold. No modification provides well-externalized conditions that, at the same time, yield satisfactory ratings of similarity to the HRIR in general and with regard to reverberation. While there is little variation between the methods regarding reverberation, the differences in general similarity are highest at conditions with poor externalization, where the DRR manipulation is preferred over the other two methods. The differences in similarity regarding sound color for externalized conditions are more distinct. Again, the DRR method is preferred while truncation received the lowest ratings. This trend is similar for naturalness.

Informal listening lead us to the impression that truncated BRIRs of medium and short lengths stand out from the other methods with timbral colorations, most likely caused by comb filters due to interference of the early reflections with the direct sound. This may very well explain lower ratings regarding sound color and naturalness. In contrast to the modification of the reverberation time or the DRR, the early reflections are unaffected by truncation until very short lengths. Reducing the reverberation, however, may lead to a de-masking of the otherwise inaudible comb filters. Though, it should be noted that anechoic conditions, while sensitive to timbral artifacts, are rarely encountered in real life.

We investigated our hypothesis that the general rating of similarity may be decomposed into the ratings with regard to sound color and the amount of reverberation. We used multiple linear regression in order to determine the contribution of the constrained to the general ratings, as well as the interaction thereof. With y denoting the general similarity rating, and x_{sc} and x_{rev} the ratings regarding sound color and reverberation, we compared models for every possible combination of x_{sc} and x_{rev}, the interaction term $\bar{x} = \sqrt{x_{\mathrm{sc}} x_{\mathrm{rev}}}$ (the geometric mean), and an additive constant. We used the BIC [7] and R^2 as criteria for model selection. The model that simultaneously minimizes the BIC and maximizes R^2 is the the geometric mean $y \sim c \cdot \bar{x}$, with $c = 0.9$ and $R^2 = 0.69$.

5. CONCLUSION

In this contribution, we presented an experiment comparing three modification techniques regarding externalization and sound quality. We defined sound quality as the similarity to an anechoic reference signal, as we are interested in BRIRs that yield decent externalization while preserving the original sound of a recording at the same time. We showed that, for the present experiment, it is plausible to explain overall similarity by the similarity regarding sound color and reverberation as the main contributing factors.

For each method, we saw a monotonic relationship between modification depth and a decrease in externalization, associated with an increase in similarity. Beyond that, the different methods are incommensurable. A comparison between the the methods may, however, be drawn either by relating them to a common quantity, which can be either a physical measure (in our case: the temporal centroid), or another response variable recorded for the same conditions (in our case: externalization). Since the comparison via a third quantity must be interpreted with caution, we consider the latter more meaningful.

Each of the methods has its own benefits. While truncation is obviously the right choice to yield short impulse responses, it has no physical equivalent and may thus sound unnatural. It can lead to timbral artifacts which, again, to avoid is the strength of the DRR modification. To modify the reverberation time seems to be a good compromise, since it also reduces the effective length of the BRIR and may therefore be combined with truncation.

Our findings may contribute to future research in two different ways: On the one hand, they provide a foundation for the investigation of hybrid modification methods to combine, e.g., the good timbral properties of the DRR and decay method with truncation in order to yield short impulse responses. On the other hand, an analysis of the modified BRIRs regarding the binaural cues may help to further understand the mechanisms of externalization.

6. REFERENCES

[1] W. M. Hartmann and A. Wittenberg, "On the externalization of sound images," *J. Acoust. Soc. Am.*, vol. 99, no. 6, 1996.

[2] J. Catic, S. Santurette, and T. Dau, "The role of reverberation-related binaural cues in the externalization of speech," *J. Acoust. Soc. Am.*, vol. 138, no. 2, 2015.

[3] S. Li, R. Schlieper, and J. Peissig, "The effect of variation of reverberation parameters in contralateral versus ipsilateral ear signals on perceived externalization of a lateral sound source in a listening room," *J. Acoust. Soc. Am.*, vol. 144, no. 2, 2018.

[4] R. Crawford-Emery and H. Lee, "The subjective effect of BRIR length on perceived headphone sound externalization and tonal coloration," in *Audio Eng. Soc. Conv. 136*, AES, 2014.

[5] S. Werner, F. Klein, T. Mayenfels, and K. Brandenburg, "A summary on acoustic room divergence and its effect on externalization of auditory events," in *8th Int. Conf. on Quality of Multimedia Experience (QoMEX)*, 2016.

[6] H. G. Hassager, F. Gran, and T. Dau, "The role of spectral detail in the binaural transfer function on perceived externalization in a reverberant environment," *J. Acoust. Soc. Am.*, vol. 139, no. 5, 2016.

[7] A. E. Raftery, "Bayesian model selection in social research," *Sociological methodology*, vol. 25, 1995.

doi:10.25836/sasp.2019.20

SUBJECTIVE EVALUATION OF SPATIAL DISTORTIONS INDUCED BY A SOUND SOURCE SEPARATION PROCESS

Simon Fargeot[1] **Olivier Derrien**[2] **Gaëtan Parseihian**[1]
Mitsuko Aramaki[1] **Richard Kronland-Martinet**[1]
[1] Aix Marseille Univ, CNRS, PRISM, Marseille, France
[2] Université de Toulon, Aix Marseille Univ, CNRS, PRISM, Marseille, France
fargeot@prism.cnrs.fr

ABSTRACT

This paper presents a new methodology to perceptually evaluate the spatial distortions that can occur in a spatial remix context, when using sound sources degraded by a source separation process. It consists in comparing localization performances on complex sound scenes composed of synthetic musical instruments in both clean and degraded cases. The localization task takes into account perceived position or positions of each instrument, as well as their perceived size and shape. In order to deal with this complex task, the test is performed through a virtual environment, using head-mounted gear. This methodology has been tested to evaluate spatial image distortions induced by an NMF source separation algorithm developed by Simon Leglaive [1]. The present study reveals that the source separation process leads to perceptible degradations of the spatial image. Three main kinds of spatial distortions have been characterized, including "phantom" sources emergence, source widening and increasing of the localization blur.

1. INTRODUCTION

The fields of video games, simulations and virtual reality are now tending to develop increasingly high-performance, realistic and immersive technologies. Efforts are made in terms of sound devices and sound processing to synthesize realistic sound scenes in a 3-D environment [2]. One challenge is the ability to analyze a 3-D audio stream corresponding to a complex sound scene in its basic components (i.e. individual sound sources), to modify the spatial scene (e.g. to change sound sources position) and to resynthesize a modified 3-D audio stream. This situation is referred to as spatial remix. Performing a spatial remix supposes reliable source separation algorithms. Such algorithms already exist but they are not perfect: recovered source signal suffer from several distortions. Objective and subjective evaluation of separation artifacts have been conducted [3], [4] and according to Emiya [5], three types of separation artifacts have been characterized, including alteration of the target source and rejections of other sources into the target source. These criteria can be determined using two Matlab toolboxes: *Perceptual Evaluation methods for Audio Source Separation (PEASS)* [5] and *Blind Source Separation Evaluation (BSS-Eval)* toolbox [6]. However, these studies usually consider the separated source signals alone, i.e. when each source is listened to separately. This is different form the spatial remix problem, where all sources are rendered simultaneously. Liu et al. proposed a method for evaluating the quality of source separation in a spatial remix context using standard objective criterions only [7]. However as the finality of this work is intended to humans, it is important to subjectively evaluate spatial distortions induced by these separation artifacts. The main difficulty with subjective evaluations is the ability to objectivate the perception of auditors in terms of spatial soudscape, such that the results can be aggregated over a large number of subjects.

This paper aims to characterize and quantify perceptually these spatial distortions by conducting a localization test on both degraded and clean versions of the same polyphonic musical extract. Localization performances in both cases are then compared taking the clean sources case as a reference. The methodology is applied to assess the quality of Non-Negative Matrix Factorization source separation algorithm developped by Leglaive [1] which performs separation on convolutive mixtures. This algorithm is introduced in the first section. Then, the experimental design and results of this study are presented followed by a discussion.

2. AUDIO SOURCE SEPARATION

Source separation is a major research theme in signal processing. The basic principle is to estimate a number N components of a mixture from a number P of observations of this mixture. It is defined in the literature according to different criteria relating to the nature of the observed mixture: instantaneous (e.g. at the exit of a mixer) or convolutive (e.g. during a recording in a concert hall, with room effects), but also to the nature of the observations of this mixture: over-determined or under-determined as well

© Simon Fargeot, Olivier Derrien, Gaëtan Parseihian, Mitsuko Aramaki, Richard Kronland-Martinet. Licensed under a Creative Commons Attribution 4.0 International License (CC BY 4.0). **Attribution:** Simon Fargeot, Olivier Derrien, Gaëtan Parseihian, Mitsuko Aramaki, Richard Kronland-Martinet. "Subjective Evaluation of Spatial Distortions Induced by a Sound Source Separation Process", 1st EAA Spatial Audio Signal Processing Symposium, Paris, France, 2019.

doi:10.25836/sasp.2019.15

as the nature of the observed sources: stationary or time varying sources. Depending on these criteria, several separation techniques can be applied [8].

2.1 Source separation using Non-negative Matrix Factorization

For under-determined convolutive recording of a musical piece, Non-negative Matrix Factorization methods seems to be a suitable choice [9]. It relies on the analysis of the energy of the mixture in the time-frequency domain, assuming that the phase of the signal is invariant over time. This involves approximating the time-frequency matrix S_{ij} of the energy sources j by the product of a spectral component matrix $\Phi_{jk}(f)$ by a time activation matrix $e_{jk}(n)$, as follows:

$$|S_{ij}|^2 \simeq |\tilde{S}_{ij}|^2 = \sum_{k=1}^{K_j} e_{jk}(n)\Phi_{jk}(f) \qquad (1)$$

where e_{jk} and Φ_{jk} are non-negative. The k components of each source are estimated from the time-frequency representation (TFCT) of the mixture energy using its parsimonious and non-negative properties. In practice, NMF-based algorithms use an iterative process to best estimate the k components of the e and Φ matrices of each source.

2.2 Leglaive's NMF algorithm

The separation toolbox used in this study is provided by S. Leglaive, R. Badeau and G. Richard in [1]. It is developed for Matlab and requires a couple of data regarding the mixture: audio file of the mixture, number of sources to be estimated, number of observations of the mixture. In addition, this algorithm does not perform a blind source separation since it requires strong prior knowledge of the nature of the mixture. Practically, source separation is performed in "oracle" mode. It requires to initialize the algorithm with a good approximation of the mixture filters impulse responses. They contain spatial cues of the sources in relation to the measurement device as well as information on the room in which the recording is performed. In this study, we wish to remain as close as possible to real use-cases. We consider the separation of acoustic mixes. Thus, the impulse responses must be estimated a priori before running the separation algorithm.

2.3 The acoustic mixes

The acoustic mixes are composed of three sound sources distributed in space and picked up by two cardioid microphones in ORTF conguration, as shown in Fig. 1.

Concerning mixture filters IR, these correspond to the transfer functions of the acoustic channel between the sources and the microphones. Under concert conditions, the number of sound sources on stage can be high and their position can vary over time. Measuring impulse responses at any point in the scene is tedious and unrealistic. However, a measurement of impulse responses at any location can already provide us with information on the late room

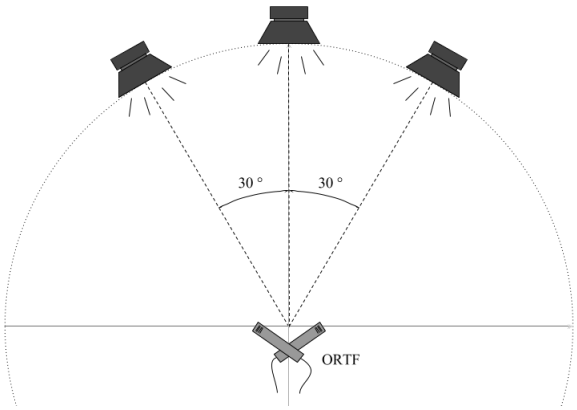


Figure 1. Apparatus for the production of real mixtures, composed by three sources and recorded by two cardioid microphones in ORTF configuration.

effect. If the positions of each source and microphones are known, it is possible with a single IR measurement to approximately build the corresponding impulse responses, by applying the right gain / delay matrix to the original IR. Assuming that they are correctly estimated, differences with the real case remain in the error of estimating the early room effects which actually depend on the position of the sources, especially at low frequencies. From the impulse responses of the central source (see Fig. 1) and by applying correct gain/delay pairs for the two peripheral sources, their impulse responses have been estimated. These fabricated responses were then used to initialize the algorithm and separation have been performed on four different musical extracts recorded in Fig. 1 conditions.

2.4 Spatial remix

In order to guide the implementation of the perceptual test, an informal listening session was conducted on the separate sources of the four mixes. The artifacts induced by the separation process are perceptible. The timbre and transients of the separate sources are slightly damaged and rejections are also present. Spatial remixes of the recovered source signals are then performed according to the spatial configuration presented in Fig. 2.

For each of the four mixes, the three recovered instruments are displayed in a "real source" format: one source per loudspeaker. This way, each mix can be rendered in three different spatial configurations obtained by changing the order of sources with respect to the loudspeaker setup. Thus, the subjective evaluation of spatial distortions induced by spatial remix of separated sources by Leglaive's NMF algorithm can be performed.

3. EXPERIMENTAL DESIGN

The purpose of this perceptual experience is to verify if the spatial image of a spatial remix performed with separated sources is perceived as significantly degraded and to characterize these distortions. According to perceptual mechanisms involved in spatial hearing [10], one can assume

doi:10.25836/sasp.2019.15

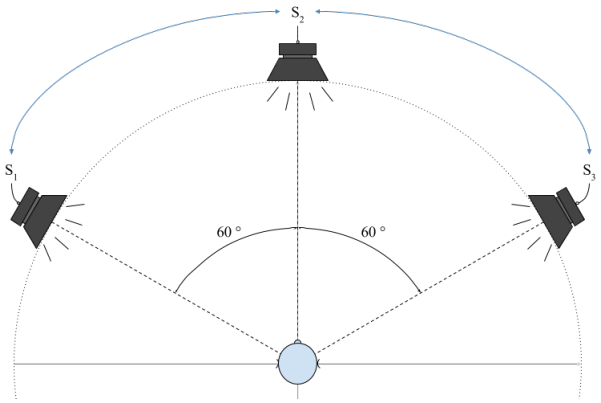


Figure 2. Spatial configuration of the sound scenes presented for the perceptual experience. The sources are placed on a circle with a radius R = 1.5m, on the azimutal plane. S_1, S_2 and S_3, are the three sources composing a sound scene. One source per speaker. The blue arrows represent the permutations of the sources corresponding to the different spatial configurations of an extract.

that distortions of the spatial image can be characterized in terms of :

- change in the perceived source position,
- change in the perceived source extent,
- unstable source position in time,
- virtual sound sources in false positions (phantom sources).

Therefore, the design of the experimental system has to allow people to report these phenomena, through a simple task.

3.1 Experimental set-up

The subject of the experiment is placed in the center of 3m diameter sphere equipped with 42 loudspeakers [11]. The spatial remix is carried out by 3 loudspeakers in the spatial configuration illustrated in Fig. 2. He or she is equipped with a virtual reality headset in which a virtual scene is projected. The visual environment was developed with Unity. It looks like a dark blue sphere with the same dimensions as the loudspeaker sphere. Using a Wiimote, the subject is able to draw in this virtual space, as shown in Fig. 3. Positions of the head of the subject and the Wiimote are tracked using the Optitrack motion catpure system.

3.2 Subjects

The population sample selected for this study consists of 20 healthy subjects (15 male and 5 female), aged between 23 and 40 years. They do not have hearing problems. 50 % of them are used to listening to music on a stereophonic or binaural spatialization device.

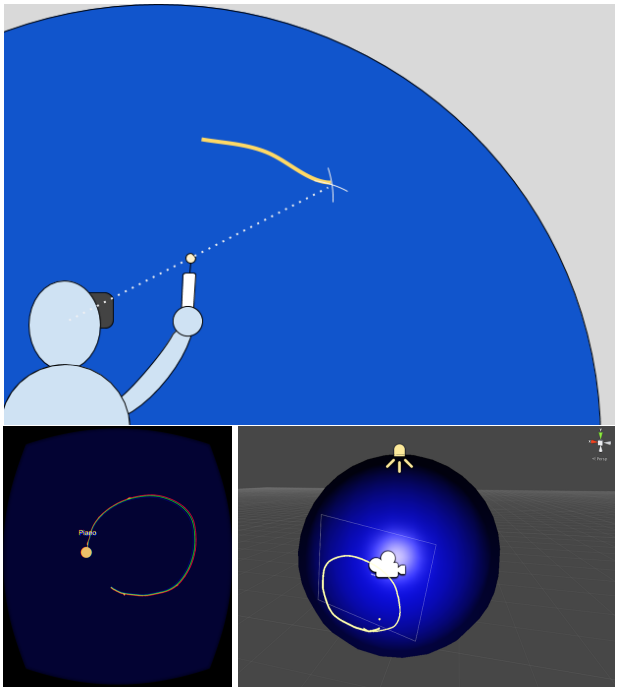


Figure 3. Virtual environment and report method. Top image represents the report method. Bottom left image is a screen-shot of the subject's view. Bottom right image is a view of the virtual environment from the outside.

3.3 Stimuli

The study is conducted on 24 sound stimuli detailed as follows:

- 4 musical extracts composed of 3 instruments,
- for each extract:
  - 2 qualities: reference quality and quality degraded by the source separation process described in section 2.3),
  - 3 spatial configurations, corresponding to the three different permutations possible with one source per loudspeaker (see: figure 2).

The musical extracts are played in a loop, without interruption, until the subject has located their 3 instruments. In the end, 8 conditions are tested (4 extracts $\times$ 2 versions). The influence of spatial arrangement is not studied here, the 3 spatial configurations of each extract are considered as 3 repetitions of the same condition.

3.4 Procedure

This test is a sound source localization test where the subject is asked to surround in a virtual environment each instrument of musical mixes of 3 instruments. It is divided into three successive stages. The first step is a demonstration phase. Five drum extracts at different positions with different apparent widths are presented with their associated plot in the virtual environment. The purpose of this step is to give the subject an overview of the different scenarios that he may encounter during the test phase.

In the second step, which is a learning phase, the subject becomes familiar with the task he or she will have to perform during the test phase. The same five drum extracts are again presented and the subject is asked to spot the source position by surrounding as precisely as possible the area in which he perceive the sound.

The third and final step is the test phase. The 24 musical extracts composed of 3 instruments (*e.g.* drums, piano, bass) are presented randomly and the subjects are asked to focus successively on each element of the mix in order to locate them. Subjects must each time define the area in which they perceive the instrument they are being asked to locate. As in the learning phase, they must surround this area using a remote pointer. The average duration of the test is 45 minutes. As the subject must remain focused throughout the test, he is advised to take breaks as soon as they experience symptoms of fatigue.

3.5 Data analysis and processing

The plots collected from the 20 subjects have more or less complex patatoid shapes and cannot be analyzed in their raw state. Data processing is necessary to retrieve statistically analyzable descriptors. Examples of drawings are provided in Appendix B. Based on the general appearance of the shapes obtained and to simplify the analysis, we consider that they can be approximated by ellipses. The analysis can thus be carried out on the different parameters of the ellipses. For each extract and for each instrument, we determine:

- the number of additional perceived sources, an indicator of the presence of phantom sources, calculated from the number of plots drawn. Each instrument is described by at least one plot, so only additional plots are counted.
- azimuth and elevation localization errors, calculated from the position of the centre of gravity of the ellipses relative to the actual position of the sources.
- the perceived width of the instrument, calculated from the product $a \times b$ of the major axis a by the minor axis b of the ellipse. In the case where several plots have been made for the same instrument, the sum of the ellipses areas is calculated.

Each descriptor is averaged over the 3 repetitions of each condition. A repeated measures ANOVA is conducted, based on the 8 test conditions. It is an ANOVA with two intra-subject factors: the type of mix (4 levels: 4 different extracts) and quality (2 levels: reference quality and degraded quality). In addition, post-hoc tests is carried out to determine the presence of interactions between factors.

4. RESULTS

The general results for all descriptors are presented in the present section. The effects of the quality and the type of mix are finally described on three descriptors as elevation localization error doesn't vary significantly across test conditions.

4.1 Effects on the number of additional sources perceived

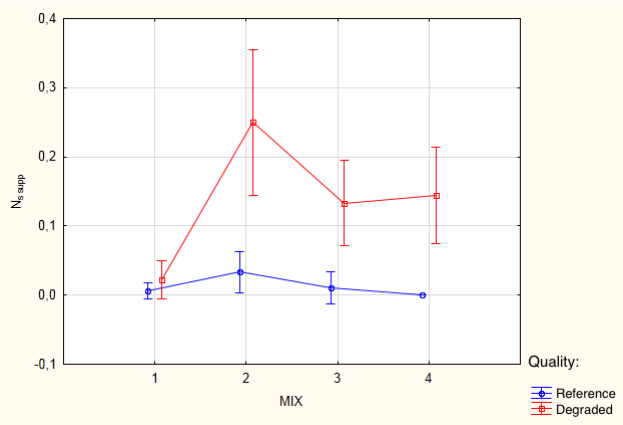


Figure 4. Cross effects of quality and mix type on the number of additional sources perceived. The red squares and blue dots represent the average value. N_{ssupp} corresponds the average of additional sources obtained for all subjects, instruments and repetitions of an extract broadcast in a given quality.

The effects of quality and mix type on the number of additional sources perceived are shown in Fig. 4. ANOVA revealed significant effects of quality ($QUAL$: $F_{1,19} = 27,643$, $p \leq 0,001$), and mix type (MIX : $F_{3,57} = 12,390$, $p \leq 0,001$) on this descriptor. The number of perceived phantom sources is significantly higher in the case of degraded scenes than in the case of reference scenes. There is also a significant interaction between quality and mix type ($F_{3.57} = 10.469$, $p \leq 0.001$). This interaction reflects the fact that the effect of quality varies greatly from one mix to another. Indeed, as shown on Fig. 4 ghost sources are heard much more frequently for mix 2 than for the others. Mix 1, on the other hand, presented almost no additional sources.

4.2 Effects on the perceived source width

The quality and type of mix also have a significant influence on the perceived width of the instruments, as shown in Fig. 5. The ANOVA gives for quality $QUAL$: $F_{1.19} = 8.4072$, $p \leq 0.01$ and for mix type MIX: $F_{3.57} = 7.1599$, $p \leq 0.005$. According to Fig. 5, degraded sources are perceived on average to be larger than intact sources. It can also be noted that the width estimation is subject to great variability. This figure also highlights the differences on the estimated area of the instruments as a function of the mix type. It appears that on average the instruments of mix 2 were perceived to be larger than those of the other extracts. Extract n1 is not significantly different from the others, however, the evaluation of the width of its sources is subject to great variations from one individual to another.

doi:10.25836/sasp.2019.15

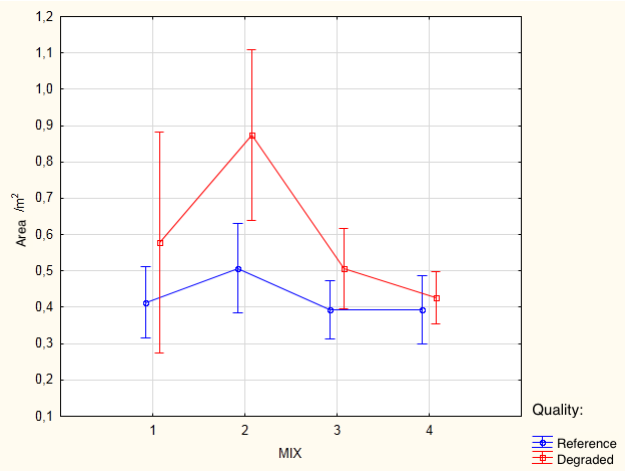


Figure 5. Cross effects of quality and mix type on perceived source width.

4.3 Effects on azimuth localization error

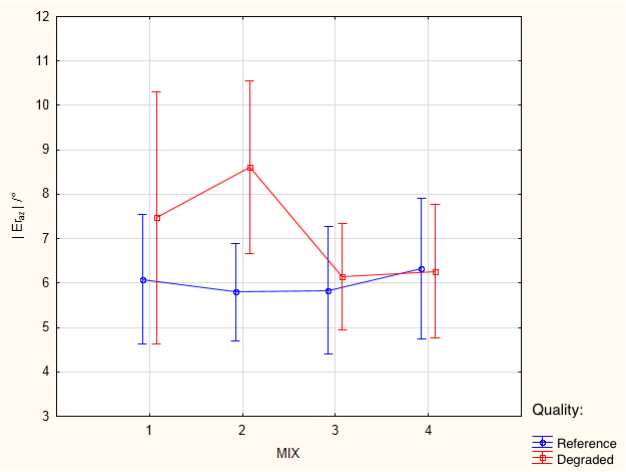


Figure 6. Cross effects of quality and mix type on azimutal localization error.

The results of the ANOVA for azimutal localization error are shown in figure Fig. 6. Only quality has a significant influence on this descriptor ($F_{1.19} = 4.9737$, $p \leq 0.05$). The azimutal localization error is on average 6 for the reference versions and 7 for the degraded versions. This difference is relatively small regarding general human localization performances in the azimutal plane. No significant variation in azimuth error is observed depending on the type of mix.

5. DISCUSSION

The results of this experiment shows perceptible differences between sound scenes composed by separated sources and clean sound scenes. First, the number of perceived phantom sources increases significantly in degraded scenes. This is due to the fact that the source separation process generates distortions such as rejections. That is, since the source estimation is not perfect, portions of the signal from one source can be attributed to the other sources of the mixture. In a spatial remix context, this means that portions of the target source are broadcast on several channels simultaneously. When the rejections are very different from the source signal in terms of temporal and spectral envelopes, the listener perceives several distinct sources [12] and in this particular case it is perceived has one main source and one or more phantom sources. Percussive instruments characterized in the time-frequency domain by short time activation and wide frequency distribution are very susceptible to rejection problems. It should be noted that phantom sources are not necessarily perceived at the position of the sources in which the rejections appear. Generally speaking, they were perceived as high and very volatile, which made their location tedious.

Second, this study reveals that the sources have been generally perceived as broader in the case of scenes degraded by source separation. The presumed reasons for this enlargement are diverse. Source rejections seem to be a plausible cause of this expansion. Indeed, Blauert [10] showed that if two coherent sources are broadcast simultaneously, they can be perceived as one large source depending on their coherence level. If rejections are coherent enough the result can be perceived as one large source. Moreover, it seems that the poor estimation of impulse responses at low frequency for the initialization of the separation algorithm has an influence on source widening since it has mainly been observed for low frequency instruments (e.g. the bass part of mix 2).

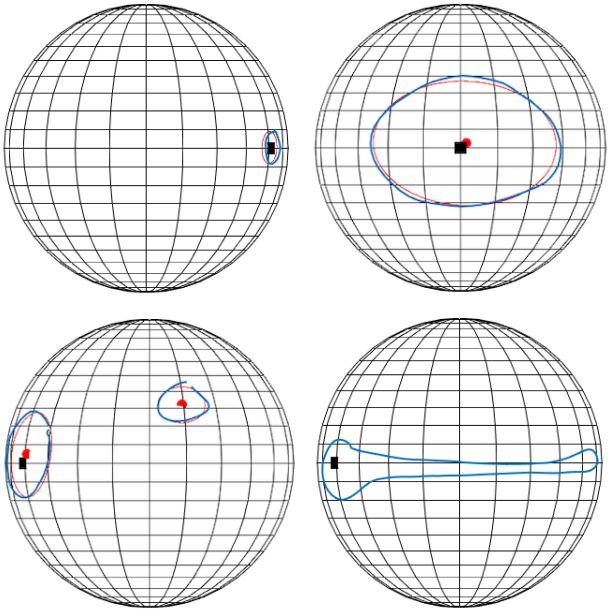


Figure 7. Examples of plots made by subjects. Black squares: real position of the source, blue curves: drawn plots, red curves: fitting ellipses, red dots: center of ellipses. From left to right and top to bottom: source perceived as punctual, source perceived as wide, additional source perceived and source perceived as unstable.

Finally, according to the results, separation artifacts

lead to a slight increase in the azimuth localization error. This is linked to enlargement and the perceived instability of position for some sources. Indeed, on the one hand when the sources are large, the estimation of the position of the source's barycentre is more imprecise and on the other hand when the sources position is unstable, it is difficult to judge a central position of the source. Some subjects have managed in their drawings to account for this phenomenon (see Fig. 7) but the elliptical approximation of these plots is no longer appropriate. In addition, the reporting method implemented proved to be effective and particularly well adapted to the task the subjects were to perform. However, the diversity of the cases presented is not fully represented by this method. For example, it does not distinguish the case of an extended source from the case of a source with an unstable position.

6. CONCLUSION

Our study revealed that a source separation process could lead to perceptible degradations in the spatial image of spatialized musical sound scenes. Three types of degradation were observed. In the majority of degraded cases, we see the emergence of " phantom " sources, an increase in the perceived width of the sources and an increase in the error made in the localization task. Percussive or harmonic rejections seems to be one cause of disturbance in the spatial image. On the other hand, the approximation of the impulse responses used to initialize the separation algorithm may have negative impact on the estimation of target sources at low-frequency. Therefore, to have an accurate restitution of the spatial image from separate sources, it would be necessary, according to this study, to minimize the error in estimating the impulse responses of the different sources and to reduce the rejection rate.

In addition, the study was conducted on subjects with different profiles and different listening skills. Some subjects were not sensitive to differences in conditions. Others, on the contrary, paid particular attention to the details of the sound scenes. It would therefore be interesting to further characterize the degradation, by running this test on an expert audience of spatialized listening. In the future, other sound examples could be studied, in particular examples that are more difficult to solve for the source separation algorithm, such as scenes composed exclusively of instruments with similar timbres (*e.g.* string quartet), percussion ensembles or environmental sounds. We chose as a first study to focus on a simple case of spatialization: with no virtual sources. A new experiment would consist in studying the performance of other spatialization techniques such as VBAP or HOA in a context of restitution of sound scenes resulting from a separation process.

7. ACKNOWLEDGMENTS

This work was made possible thanks to the help of Simon Leglaive, Roland Badeau and Gaël Richard from LTCI - Télécom ParisTech, who provided us with the audio source separation algorithm used for this study.

8. REFERENCES

[1] S. Leglaive, R. Badeau, and G. Richard, "Multichannel audio source separation: variational inference of time-frequency sources from time-domain observations," in *42nd International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017.

[2] S. Spors, H. Wierstorf, A. Raake, F. Melchior, M. Frank, and F. Zotter, "Spatial sound with loudspeakers and its perception: A review of the current state," *Proceedings of the IEEE*, vol. 101, no. 9, pp. 1920–1938, 2013.

[3] J. Kornycky, B. Gunel, and A. Kondoz, "Comparison of subjective and objective evaluation methods for audio source separation," in *Proceedings of Meetings on Acoustics 155ASA*, vol. 4, p. 050001, ASA, 2008.

[4] B. Fox, A. Sabin, B. Pardo, and A. Zopf, "Modeling perceptual similarity of audio signals for blind source separation evaluation," in *International Conference on Independent Component Analysis and Signal Separation*, pp. 454–461, Springer, 2007.

[5] V. Emiya, E. Vincent, N. Harlander, and V. Hohmann, "Subjective and objective quality assessment of audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2046–2057, 2011.

[6] C. Févotte, R. Gribonval, and E. Vincent, "Bss_eval toolbox user guide–revision 2.0," 2005.

[7] Q. Liu, W. Wang, P. J. Jackson, and T. J. Cox, "A source separation evaluation method in object-based spatial audio," in *2015 23rd European Signal Processing Conference (EUSIPCO)*, pp. 1088–1092, IEEE, 2015.

[8] E. Vincent, M. G. Jafari, S. A. Abdallah, M. D. Plumbley, and M. E. Davies, "Blind audio source separation," *Queen Mary, University of London, Tech Report C4DM-TR-05-01*, 2005.

[9] A. Ozerov and C. Févotte, "Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 3, pp. 550–563, 2010.

[10] J. Blauert, *Spatial hearing: the psychophysics of human sound localization*. MIT press, 1997.

[11] G. Parsehian, L. Gandemer, C. Bourdin, and R. K. Martinet, "Design and perceptual evaluation of a fully immersive three-dimensional sound spatialization system," in *3rd International Conference on Spatial Audio (ICSA 2015)*, 2015.

[12] A. S. Bregman, *Auditory scene analysis: The perceptual organization of sound*. MIT press, 1994.

doi:10.25836/sasp.2019.15

EEG MEASUREMENT OF BINAURAL SOUND IMMERSION

Rozenn Nicol[1] **Olivier Dufor**[2] **Laetitia Gros**[1]
Pascal Rueff[3] **Nicolas Farrugia**[2]
[1] Orange Labs, 2 avenue Pierre Marzin, 22300 Lannion, France
[2] CNRS Lab-STICC, Technopole Brest-Iroise, 29238 Brest, France
[3] Agence du Verbe, 2 Keraudren, 22260 Plouec Du Trieux, France
rozenn.nicol@orange.com

ABSTRACT

The purpose of this study is to examine whether neuroimaging could provide new insights into the assessment of the listening experience of spatial audio content. Our objective is to explore cognitive processes involved when listening to spatial audio content by electroencephalography (EEG) measurement. The experiment is based on a reversed implementation of the oddball paradigm. Audio stimuli are used to interfere with a detection task of visual stimuli. It is expected that the distracting effect of binaural reproduction is stronger than that of stereophony. Evidence of this is sought in both behavioral data (i.e. the performances in detecting deviant stimuli) and EEG-HR measurements (e.g. P300 response). The differential response between standard and deviant stimuli was examined as a function of the audio distractor. It was observed that the P300 amplitude was significantly higher in presence of the binaural distractor than in the stereophonic condition, suggesting a stronger effect of surprise possibly caused by a more immersive reproduction. This was confirmed by behavioral results which showed a longer response time for binaural distractors (566 ms) than for stereophonic ones (550 ms).

1. INTRODUCTION

The purpose of this study is to examine whether neuroimaging could provide new insights into the assessment of the listening experience of spatial audio content. Conventional methods of assessing sound reproduction are based on either Quality of Experience (QoE) scores or localization judgments [1]. Their main disadvantage is that the subject is aware of his (her) rating task. Besides, the assessment is restricted to specific dimensions (i.e. perception of degradation or spatial attributes). Dimensions such as emotions or cognitive load are generally not taken into account. Therefore, our objective is to explore cognitive processes involved when listening to spatial audio content by EEG (Electroencephalography) measurement. Descriptors of the listening experience (e.g. immersion, realism) are sought in the electrical activity of the brain.

Neurophysiological methods are already identified as a promising way of investigating the field of QoE assessment [2, 3]. A study showed that some EEG features are correlated with emotion primitives (i.e. valence and arousal) [4]. They were successfully used to predict the influence of human factors (i.e. users' perception, emotional and mental state) on a QoE score for a comparison of text-to-speech and natural speech. In a similar study, preference judgments were related to functional near-infrared spectroscopy (fNRIS) features [5]. However, so far there have only been a few experiments that implemented these methods to assess spatial-audio technologies. A first one compared spatial-response fields of the primary auditory cortex with virtual sound sources synthesized with individual and non-individual HRTFs in ferrets [6]. It was shown that the responses obtained with an animal's own ears differed significantly in shape and position from those obtained with another ferret morphology. More recent work has confirmed a positive correlation between various levels of accuracy of spatial sound reproduction (e.g., individual HRTF vs. generic HRTF vs impoverished localization cues [7,8], natural vs. artificial cues of auditory motion [9], individual binaural vs. stereo recordings for sound externalization [10]) and the activity of the auditory cortex measured either by magnetoencephalography (MEG), by EEG or by fMRI. Most of the results reported above were obtained by using binaural synthesis.

In the present study, the quality of spatial sound reproduction will be assessed in terms of its capacity to distract the subject from a task of visual detection. Binaural and stereophonic reproduction will be compared. It is expected that the distracting effect of binaural reproduction is stronger than that of stereophony. Evidence of this will be sought in both behavioral data (i.e. the performances in detecting deviant stimuli) and EEG-HR measurements (256 sensors). First, the experiment is described. Then, all the preprocessing of EEG data is detailed, before presenting results.

© Rozenn Nicol, Olivier Dufor, Laetitia Gros, Pascal Rueff, Nicolas Farrugia. Licensed under a Creative Commons Attribution 4.0 International License (CC BY 4.0). **Attribution:** Rozenn Nicol, Olivier Dufor, Laetitia Gros, Pascal Rueff, Nicolas Farrugia. " EEG measurement of binaural sound immersion", 1st EAA Spatial Audio Signal Processing Symposium, Paris, France, 2019.

doi:10.25836/sasp.2019.21



Figure 1. EGI sensor net composed of 256 electrodes.

2. METHOD

2.1 Participants

Nineteen healthy adult participants reporting normal hearing and vision participated in the experiment (9 males and 10 females, mean age = 34.3 years; min/max = 18/59 years). All participants were correctly informed of the experiment and signed a consent form before being included following the international Helsinki recommendations on human research.

2.2 Experimental setup

Subjects were seated in a comfortable chair in an electrically shielded room at the Pontchaillou Hospital in Rennes. The chair faced a computer screen (1 m distant) and was surrounded with opaque curtains, so that the participants could not see what was happening in the room.

A EGI sensor net of 256 electrodes (Fig. 1) was placed on the participant's head to record his (her) brain activity. In addition, a headphone (Sennheiser HD650) was arranged over it to reproduce sound scenes that were previously recorded in the same experimental room. The pretext for justifying the use of headphones was a listening session as a further step in the experiment. The participant was thus not aware that virtual sounds were going to be played by headphones, and that his (her) listening experience was under study. Besides, at no time was he (she) asked direct questions about the quality of sound reproduction. Interviewed after the experiment, most participants reported that they believed that the sounds really emanated from the experimenter, not from the headphone reproduction.

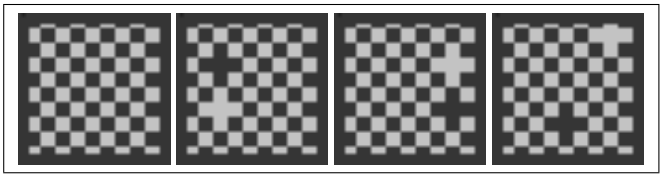


Figure 2. Visual stimuli used in the oddball paradigm: standard stimulus (first on the right) and 3 examples of deviant stimuli.

2.3 Oddball paradigm

The experiment consisted in a visual oddball paradigm programmed with E-prime software. Both accuracy and reaction time were recorded. The participant was presented a sequence of images representing small (1 cm sided) grey checkerboards randomly distributed (n=320). Standard stimuli (n=265) represented perfect checkerboards, while deviant stimuli (n=55) were representing altered checkerboards (see Fig. 2). The rare stimuli (i.e. altered checkerboards), which represented 17% of the whole set of stimuli, were introduced to generate an event-related potential (ERP), known as the P300 in the brain. Four sequences of pseudo-randomly distributed stimuli were created. The participant was asked to click whenever he (she) detects a deviant stimulus. Audio stimuli were used to interfere with the visual oddball task. The objective was to observe the effect of such distractors on the P300 response as a function of the type of sound spatialization. Every participant had to perform the task twice; with stereo and binaural sounds separately (see section 2.4). Sounds and visual stimuli were not time locked, but both stereo and binaural conditions were associated with perfectly counterbalanced sequences of visual stimuli across subjects.

A trial was composed of the following sequence: first, a grey screen (i.e. no visual stimulus) for a variable period of time chosen to be either 150 or 650 ms; second, the visual stimulus for a period of 250 ms; third, a grey screen for a period of 750 ms. Thus, the participant had 1 s (250 + 750 ms) from the moment the stimulus was displayed, to click if he (she) detected a deviant stimulus. At the beginning of the experiment, the participant carried out a training phase for the visual detection task. At the end, he (she) had to complete a questionnaire about the task difficulty. It was also checked whether he (she) realized that sounds were coming from headphones. Triggers corresponding to the onsets of visual stimuli were directly sent to the EEG recording system.

2.4 Audio stimuli

The sound scenes contained ambient noise in combination with isolated sounds potentially emanating from the experimenter behind the opaque curtains (e.g. falling keys, running water from a tap, leafing through a book, quietly walking around), but they were totally free from speech and music signals. Sounds were delivered thanks to a Focusrite Scarlett 6i6 external sound interface connected to a Mac mini computer. The listening level was adjusted to

doi:10.25836/sasp.2019.21



Figure 3. Recording setup.

60 dBA. The oddball experiment was repeated twice: once with the binaural recording (by a dummy head Neumann KU 100, see Fig. 3) and the other time with the stereophonic recording (by a XY pair, see Fig. 3). The presentation order was randomized. Intensity stereophony (i.e. XY recording) was chosen to provide the highest contrast with binaural spatialization which uses both interaural differences of time and level, as well as spectral cues. The binaural and stereophonic scenes were similar but not identical to prevent any habituation effect.

2.5 Preprocessing of EEG data

EEG signals were recorded with NetStation software and saved as raw files with triggers. They were then preprocessed using custom scripts based on Fieldtrip [11] (http://fieldtriptoolbox.org) and MNE-Python [12]. Preprocessing followed a predefined pipeline consisting of the following steps.

i) Channel removal (channels for which experimenters noticed high impedance values during subjects recordings, mostly in the neck and on the face of participants) and raw signal inspection, annotation and filtering (0.5Hz-45Hz).

ii) Isolating signals of interest around trials onsets (from -200ms to 800ms) and semi-automatic artefact annotations (jumps and muscle activity).

iii) Visual inspection of annotated artifacts, channels to be interpolated and trials to eliminate.

iv) Independent component analysis (ICA) on raw trials with annotated time segments and channels from steps i, ii and iii.

v) Visual inspection of components (the sixty first ones) and removal of components sharing blinks, saccades and cardiac activity mostly.

vi) Back-projection of left components in the channel space for signal reconstruction, removal of annotated artefacts or trials and interpolation of annotated channels.

vii) Calculation of event related potentials (ERP) and statistical analysis.

At the end of the aforementioned preprocessing steps, we obtain 124 electrodes left across all participants. The data set consisted of 6080 trials, 1045 of which were deviants. ERPs were averaged per electrode and per subject for each condition. The differential response between deviant and standard stimuli was examined as a function of the audio distractor (binaural or stereophonic).

2.6 Analysis of EEG data

We first ran a cluster analysis (number of permutations = 1000) using Fieldtrip [11], to identify groups of adjacent electrodes showing a P300 effect in both stereo and binaural conditions. We then used these channels to build regions of interest (ROI) wherein we looked for differences in the P300 amplitude for stereophonic vs. binaural sounds. More exactly, the differential response between deviant and standard stimuli is compared between the stereophonic and binaural conditions (i.e. $[\text{Deviant} - \text{Standard}]_{\textbf{binaural}} - [\text{Deviant} - \text{Standard}]_{\textbf{stereophonic}}$). ROI were made of equal numbers of channels (n=17). The P300 amplitude difference was analyzed on a 20 ms sliding window from 280 ms to 440 ms with a 10 ms overlap.

3. RESULTS

3.1 Behavioral results

Our paradigm integrates sounds as interference disturbing the main task. The literature has already reported that binaural sounds are more immersive than stereophonic sounds [13]. Consequently, our hypothesis was that accuracy scores and reaction times would be respectively decreased and increased in the binaural condition compare to the stereophonic condition. Accordingly, we used unilateral Student paired t-test to report behavioral results. In terms of accuracy, we did not found any difference between the two listening conditions (binaural vs. stereophonic) with p=0.46 for missed deviants and p=0.064 for

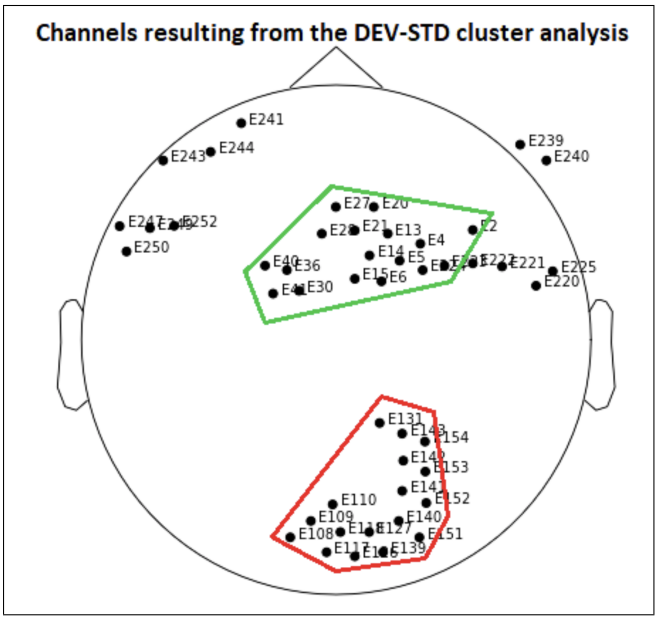


Figure 4. Channels exhibiting a P300 response resulting from the cluster analysis of the contrast DEVIANT-STANDARDS for both conditions (stereophonic and binaural sounds).

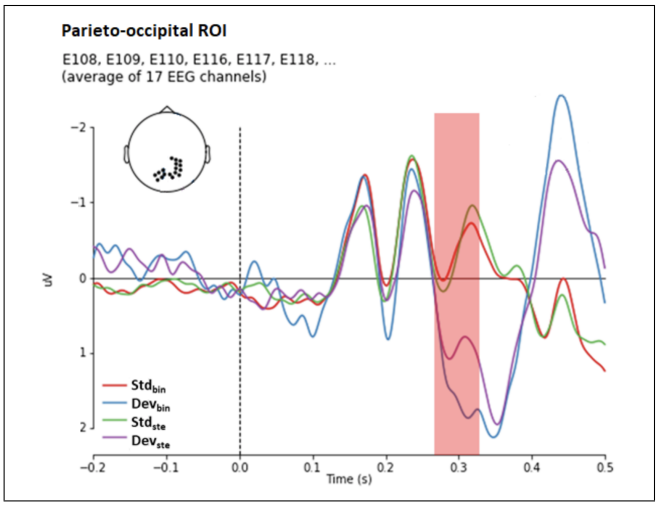


Figure 5. ERPs averaged over the 17 channels of the parieto-occipital region. Blue and purple curves correspond to $\text{Deviant}_{\textbf{binaural}}$ and $\text{Deviant}_{\textbf{stereo}}$ respectively, while red and green curves correspond to $\text{Standard}_{\textbf{binaural}}$ and $\text{Standard}_{\textbf{stereo}}$ respectively. The blue and purple curves differ significantly over the time period highlighted in pink (from 280 ms to 319 ms).

false recognitions. However, reaction times (RT) were significantly longer for the binaural condition than for the stereophonic condition with p=0.027 (binaural mean RT = 566 ms and stereophonic mean RT = 550 ms). The fastest participant used to press the response button on average 440 ms after the image onset. His average RT was chosen to stop the signal analysis and prevent any interpretation of signals after this delay.

3.2 EEG results

3.2.1 Cluster analysis

Inspection of individual ERPs revealed that the P300 was not starting before 280 ms in our dataset. Testing for a P300 effect in the latency range from 280 to 440 ms post-stimulus, the cluster-based permutation test showed a significant difference between the deviant and standard stimuli, either with binaural or stereophonic interferences ($p<0.05$). In this latency range, the difference was most pronounced over two central groups of channels (frontal and parieto-occipital). For the binaural condition, the analysis indicated the presence of 31 significant clusters within this time period while only 18 significant clusters were found for the stereophonic condition. The channels that form clusters in both conditions are presented on Fig. 4. We created two ROI from this channel set to run the amplitude comparison analysis. The first ROI groups 17 channels in the frontal part of the brain (area delimited by a green line on Fig. 4) and the second ROI groups 17 channels in the parieto-occipital region (area delimited by a red line on Fig. 4).

3.2.2 Compared analysis of P300 amplitude

The comparison of P300 amplitude consists in investigating whether deviant visual stimuli generate larger amplitude of the P300, or part of it, in a given condition compared to the other (binaural or stereophonic). We observed that only the binaural condition was associated with larger P300 in both ROI. The effect was precocious for the parieto-occipital ROI and was immediately followed by the frontal ROI. After 350 ms, deviants did not produce different P300 signals. Looking at ERP curves in each ROI (Fig. 5 and Fig. 6) and statistical data of Tab. 1, we identified that only the early phase of the P300 is concerned. This phase corresponds to the P3a. The second phase of the P300 - clearly observable on Fig. 5 (second peak) - seems to correspond to the P3b. This phase was not affected by sound interferences.

4. DISCUSSION

Taken together our results are in favor of a better surprise effect mediated by binaural sounds when compared to stereophonic sounds. We observed that the P300 amplitude was significantly higher in presence of the binaural distractor especially during its first phase known as the P3a [14]. This part of the P300 is described as reflecting surprise or novelty among sensory inputs or at least transitory losses of attentional focus when competing tasks interfere. Associated with longer reaction times, an increase of the P3a has already been described accounting for attentional switching even using multimodality [15,16]. In these studies, authors aimed at quantifying the physical similitudes or differences between relevant and non-relevant stimuli (Go-NoGo task and oddball tasks) to see which amount of them is needed to affect the electrophysi-

doi:10.25836/sasp.2019.21

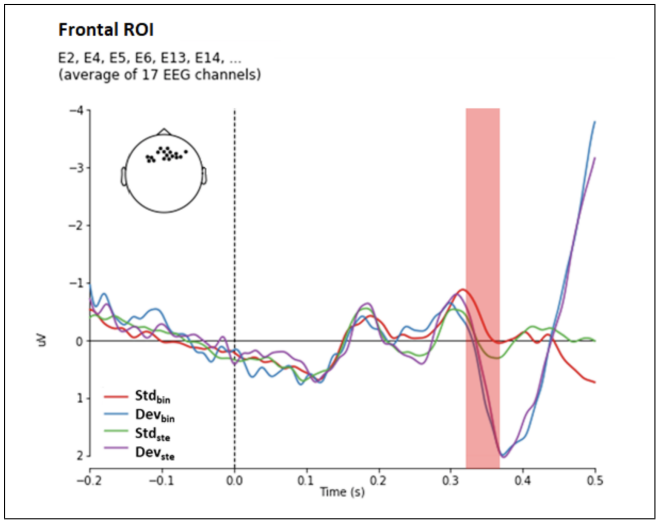


Figure 6. ERPs averaged over the 17 channels of the frontal region. Blue and purple curves correspond to $\text{Deviant}_{\textbf{binaural}}$ and $\text{Deviant}_{\textbf{stereo}}$ respectively while red and green curves correspond to $\text{Standard}_{\textbf{binaural}}$ and $\text{Standard}_{\textbf{stereo}}$ respectively. In the time period where the difference between the ERP of deviant stimuli is significant and which is highlighted in pink (from 320 ms to 359 ms), the difference between the red and blue curves is slightly larger than the difference between the purple and green curves, suggesting that deviant stimuli were affected in the binaural condition and not in the stereophonic one.

Time period (ms)	Parieto-occipital ROI	Fronto-central ROI
[280-299]	T= 2.52, p=0.021*	T=-0.21, p=0.833
[290-309]	T=2.64, p=0.017*	T=-0.50, p=0.621
[300-319]	T=2.74, p=0.013*	T=-1.27, p=0.222
[310-329]	T=2.22, p=0.039*	T=-2.07, p=0.053
[320-339]	T=1.43, p=0.170	T=-3.04, p=0.007*
[330-349]	T=0.77, p=0.453	T=-3.41, p=0.003*
[340-359]	T=0.34, p=0.740	T=-2.74, p=0.014*
[350-369]	T=0.13, p=0.901	T=-1.72, p=0.103
[360-379]	T=0.13, p=0.898	T=-0.81, p=0.429
[370-389]	T=-0.02, p=0.985	T=-0.39, p=0.697
[380-399]	T=-0.57, p=0.578	T=-0.32, p=0.755
[390-409]	T=-1.02, p=0.323	T=-0.63, p=0.536
[400-419]	T=-1.08, p=0.296	T=-0.40, p=0.690
[410-429]	T=-1.13, p=0.272	T=0.21, p=0.840
[420-439]	T=-1.32, p=0.204	T=0.24, p=0.812

Table 1. Detailed analysis of the contrast $[\text{Deviant} - \text{Standard}]_{\textbf{binaural}} - [\text{Deviant} - \text{Standard}]_{\textbf{stereophonic}}$. The P300 period is divided into 20 ms time windows overlapping by 10 ms.

ological response. Indeed, Comerchero and Polich explain that the effect of target and non-target stimuli considering task difficulty should be considered [17]. Especially, it is shown that P3a is increased if the target/standard discrimination is difficult while the non-target/standard stimulus difference is large. For example, Schröger and Wolff, who used only audio stimuli in a three stimulus oddball paradigm, showed that task-irrelevant frequency deviants elicited MMN, N2b, and P3a components, and caused impoverished behavioral performance to targets [18]. Distractors of our experiment, and more generally in interference task, play the exact same role as the irrelevant targets in the three stimulus oddball paradigm introduced by [19]. When using multimodality, the distance between deviants and distractors is such that the only possible interpretation would be a loss of attentional focus (see [18] for a discussion between memory and attentional theory involvement given distractor and deviant similarities). In our task, the only difference between sounds lies in their recording qualities. Our results suggest that observed electrophysiological and behavioral differences are a consequence of recording qualities. We deliberately chose to use sounds recorded from the experimental room placing ourselves in the context of hyper-realism as one of the best inducer for immersion.

5. CONCLUSION

Binaural and stereophonic reproduction were compared by neuroimagery according to an oddball paradigm. It was shown that the P300 amplitude was significantly affected when the audio distractor is binaural. A similar tendency was observed on the reaction time of visual detection, which was longer in presence of binaural sound. Our study introduced a promising paradigm to objectively assess immersive properties in audio media.

6. ACKNOWLEDGMENTS

This study was co-funded by the LabEx Comin-Labs under the Neural Communications project, and by Orange. We also wish to thank all the participants that undertook the study, as well as the staff from the epileptology unit at the University Hospital.

7. REFERENCES

[1] R. Nicol, L. Gros, C. Colomes, M. Noisternig, O. Warusfel, H. Bahu, B. Katz, and L. Simon, "A roadmap for assessing the quality of experience of 3d audio binaural rendering," 2014.

[2] Z. Akhtar and T. H. Falk, "Audio-visual multimedia quality assessment: a comprehensive survey," *IEEE Access*, vol. 5, pp. 21090–21117, 2017.

[3] K. u. R. Laghari, R. Gupta, S. Arndt, J. N. Antons, R. Schleicher, S. Møller, and T. H. Falk, "Neurophysiological experimental facility for Quality of Experience (QoE) assessment," in *2013 IFIP/IEEE International*

Symposium on Integrated Network Management (IM 2013), (Ghent, Belgium), pp. 1300–1305, May 2013.

[4] R. Gupta, K. Laghari, H. Banville, and T. H. Falk, "Using affective brain-computer interfaces to characterize human influential factors for speech quality-of-experience perception modelling," *Human-centric Computing and Information Sciences*, vol. 6, Dec. 2016.

[5] K. u. R. Laghari, R. Gupta, S. Arndt, J. N. Antons, S. Møllery, and T. H. Falk, "Characterization of human emotions and preferences for text-to-speech systems using multimodal neuroimaging methods," in *2014 IEEE 27th Canadian Conference on Electrical and Computer Engineering (CCECE)*, (Toronto, ON, Canada), pp. 1–5, May 2014.

[6] T. D. Mrsic-Flogel, A. J. King, R. L. Jenison, and J. W. Schnupp, "Listening through different ears alters spatial response fields in ferret primary auditory cortex," *Journal of Neurophysiology*, vol. 86, pp. 1043–1046, Aug. 2001.

[7] K. J. Palomäki, H. Tiitinen, V. Mäkinen, P. J. C. May, and P. Alku, "Spatial processing in human auditory cortex: the effects of 3D, ITD, and ILD stimulation techniques," *Brain Research. Cognitive Brain Research*, vol. 24, pp. 364–379, Aug. 2005.

[8] M. G. Wisniewski, G. D. Romigh, S. M. Kenzig, N. Iyer, B. D. Simpson, E. R. Thompson, and C. D. Rothwell, "Enhanced auditory spatial performance using individualized head-related transfer functions: An event-related potential study," *The Journal of the Acoustical Society of America*, vol. 140, pp. EL539–EL544, Dec. 2016.

[9] S. Getzmann and J. Lewald, "Effects of natural versus artificial spatial cues on electrophysiological correlates of auditory motion," *Hearing Research*, vol. 259, no. 1, pp. 44 – 54, 2010.

[10] A. Callan, D. E. Callan, and H. Ando, "Neural correlates of sound externalization," *NeuroImage*, vol. 66, pp. 22–27, Feb. 2013.

[11] R. Oostenveld, P. Fries, E. Maris, and J. M. Schoffelen, "Fieldtrip: open source software for advanced analysis of meg, eeg, and invasive electrophysiological data," *Computational intelligence and neuroscience*, vol. 1, 2011.

[12] A. Gramfort, M. Luessi, E. Larson, D. Engemann, D. Strohmeier, C. Brodbeck, R. Goj, M. Jas, T. Brooks, L. Parkkonen, and M. Hämäläinen, "Meg and eeg data analysis with mne-python," *Frontiers in Neuroscience*, vol. 7, 2013.

[13] R. Nicol, L. Gros, C. Colomes, E. Roncière, and J.-C. Messonnier, "Etude comparative du rendu de différentes techniques de prise de son spatialisée aprs binauralisation," in *Congrès Français d'Acoustique*, SFA, 2016.

[14] J. Polich, "Updating p300: an integrative theory of p3a and p3b," *Clinical neurophysiology*, vol. 118, no. 10, pp. 2128–2148, 2007.

[15] K. Ahlo, C. Escera, R. Díaz, E. Yago, and J. Serra, "Effects of involuntary auditory attention on visual task performance and brain activity," *NeuroReport*, vol. 8, p. 32333237, 1997.

[16] J. I. Katayama and J. Polich, "Auditory and visual p300 topography from a 3 stimulus paradigm," *Clinical Neurophysiology*, vol. 110, no. 3, pp. 463–468, 1999.

[17] M. D Comerchero and J. Polich, "P3a, perceptual distinctiveness, and stimulus modality," *Cognitive Brain Research*, vol. 7, no. 1, pp. 41–48, 1998.

[18] E. Schröger and C. Wolff, "Behavioral and electrophysiological effects of task-irrelevant sound change: A new distraction paradigm," *Cognitive brain research*, vol. 7, no. 1, pp. 71–87, 1998.

[19] J. I. Katayama and J. Polich, "P300 from one-, two-, and three-stimulus auditory paradigms," *International Journal of Psychophysiology*, vol. 23, no. 1-2, pp. 33–40, 1996.

doi:10.25836/sasp.2019.21

BINAURAL SOUND RENDERING IMPROVES IMMERSION IN A DAILY USAGE OF A SMARTPHONE VIDEO GAME

Julian Moreira[1] **Laetitia Gros**[2] **Rozenn Nicol**[2]
Isabelle Viaud-Delmon[3] **Cécile Le Prado**[1] **Stéphane Natkin**[1]
[1] Cnam (CEDRIC), Paris, France
[2] Orange Labs, Lannion, France
[3] CNRS, Ircam, Sorbonne Université, Ministère de la Culture, Paris, France
Correspondance: julian.moreira.fr@gmail.com

ABSTRACT

Binaural rendering is a technology that could be advantageously coupled with a smartphone application, but is still not commonly used. The aim of this study is to investigate how this technology could enrich the experience of a video game application delivered on a smartphone, in terms of immersion, memorization and performance. We have used a longitudinal research procedure, the Experience Sampling Method, asking individuals to accomplish short game sessions in their daily life. With this procedure, we want to determine if a significant effect of binaural rendering can be detected while data are noised by realistic contextual variations. The results indicate a better feeling of immersion during the binaural sessions, but no improvement was shown for memorization and performance. However, as the experiment is deployed in realistic contexts of use, this result justifies the implementation of a binaural rendering in smartphone applications.

1. INTRODUCTION

Binaural synthesis is a technology that spatializes sound outside of the head of a listener wearing headphones. It allows to reproduce the acoustic properties of a sound coming from a specific direction by using the Head Related Transfer Functions (HRTF) filters owned by this listener [1]. Being the only solution of spatialized sound to be easily transportable everywhere, it particularly fits with the use of a smartphone, where contexts are numerous, dynamic and unpredictible.

Most of the works evaluating binaural technologies have been dedicated to measure the accuracy of source localization (see for instance [2–5]). However, in a daily life, audiovisual experiences with binaural sounds (video games, movies, videoconferencing, etc.) are not necessarily to be limited to a localization task. As such, a few other studies have focused their attention to the contribution of a binaural rendering (compared to mono or stereo renderings, or a purely visual scene, etc.) by looking at various other attributes: immersion (e.g., [6–8]), navigation performance (e.g., [6, 7]), memorization (e.g., [6]), feeling of spatialisation (e.g., [9]), consistence or correlation with visual (e.g., [9, 10]), global appreciation (e.g., [10]), etc.

To our knowledge, none of these works have been dedicated so far to a binaural rendering coupled with a smartphone application. Yet, smartphone usages raise additional questions related to contextual factors, that might influence the perception of a spatialized sound scene (see for instance [12–17] about the role of some of these factors on binaural-enhanced experiences). Other questions rely on methodology: contexts bring noise, whether it be auditory, visual or even cognitive (e.g., walking and using a smartphone at the same time), that can be hardly reproduced in a laboratory environment. In this paper, we propose an experiment to study the contribution of a binaural rendering to a smartphone video game application, in terms of immersion, memorization and performance. We address the question of influence factors by deploying our experiment in real life situations. By doing this, we want to determine if a significant effect of binaural rendering can be detected while data are noised by realistic contextual variations.

In Section 2, we introduce the notion of data validity and the deployment method we choose. In Section 3 we present our experiment, including our video game, the data we collect and the setup. Section 4 is dedicated to the results, discussed in Section 5.

2. DATA VALIDITY AND DEPLOYMENT METHOD

Data validity and deployment method are closely linked. In [18], Scriven explains that collected data are valid if they are likely to answer the question initially asked. He distinguishes two types of validity: internal and external. Internal validity is linked to the control experimenters have on influence factors, like environmental conditions (location, temperature, luminosity, time of the day, ambiant noise, etc.), expertise of the assessors, or similarity of the conditions when presenting stimuli (order, number of occurences, etc.) The more under control these factors are,

© Julian Moreira, Laetitia Gros, Rozenn Nicol, Isabelle Viaud-Delmon, Cécile Le Prado, Stéphane Natkin. Licensed under a Creative Commons Attribution 4.0 International License (CC BY 4.0). **Attribution:** Julian Moreira, Laetitia Gros, Rozenn Nicol, Isabelle Viaud-Delmon, Cécile Le Prado, Stéphane Natkin. "Binaural sound rendering improves immersion in a daily usage of a smartphone video game", 1st EAA Spatial Audio Signal Processing Symposium, Paris, France, 2019.

doi:10.25836/sasp.2019.23

the more internally valid the data become. External validity represents how close to a real use case the experiment is. As such, if a laboratory represents an experimental place with a high internal validity, on the contrary external validity is potentially low, especially when it comes to reproduce smartphone usage situations. For this reason, we propose a protocole oriented to external validity, where our experimental sessions are deployed outside the laboratory.

We use the Experience Sampling Method (ESM). This method consists in dividing a long experiment into small sessions over a long period of time, in order to collect data that represent the subjects' daily life [19]. Several times per day, they are notified to accomplish a session that should not exceed a few minutes. The short duration is intended to prevent the experiment to be intrusive, and to keep the motivation of the subjects. With this process, the experience is designed to last one or several weeks. The ESM has already been used a few times with mobile phones, for instance in [20,21]. Advantageously, the whole experiment can be processed on the device: notifications, application usages, additional questionaries, etc. Furthermore, contextual data can be collected, by extracting automatically information from sensors (which requires a solid theoretical model to interpret them) or by directly asking subjects. This is of particular interest to observe trends about daily life usages, and to see if experimental data are consistent with their context.

3. EXPERIMENT

3.1 Progress of the experiment

In our experiment, we use the ESM as follows: during five weeks, subjects are notified twice a day by SMS to play a five minutes session (for a total of 70 sessions per subject). In order to diversify contexts, one session is randomly scheduled in the morning between 8am and 1pm, the other is randomly scheduled in the afternoon between 1pm and 6pm (see Table 1). To maximize the external validity of the data, the experiment takes place on the personal smartphones and headphones of the subjects.

Each session is divided into four phases, all embedded in the same application: the context questionary, the calibration of sound volume, the game itself and the end questionary. The data gathered during all these steps are sent at the end of the session to a database located on a distant server. If a subject is not connected to the internet at this time, the data are stored locally and added for sending to those of the next session.

3.2 Contextual data collection and sound calibration

Contextual data are collected via a preliminary questionary at each session. In order to keep the session short, we limit the number of questions to four, retrieving the current location of the subjects, their social surrounding, their level of mobility and their level of occupation. The purpose of this questionary is to observe trends about smartphone contexts, and because they are not controlled factors (not equally distributed between subjects), they are not meant to be used to intepret the contribution of binaural sounds. As such, they are not in the scope of this paper, and not treated in the next sections.

The sound calibration step consists in a panel that reminds the subjects to plug their headphones, put the sound volume of the phone to its maximum value, and to calibrate the volume of the application via a slider in-app. This helps us to control sound volumes and prevent subjects from muting sound.

3.3 The video game

Our smartphone application is an Infinite Runner developped for the occasion with the game engine Unity (see Figure 1). The player controls an avatar that walks automatically on a procedurally generated road, and with an increasing speed. The goal is to bring the avatar as far as possible, avoiding obstacles and gathering bonus items. To do that, the player can swipe a finger on screen to move the avatar on the right, on the left, make him jump or slide down.

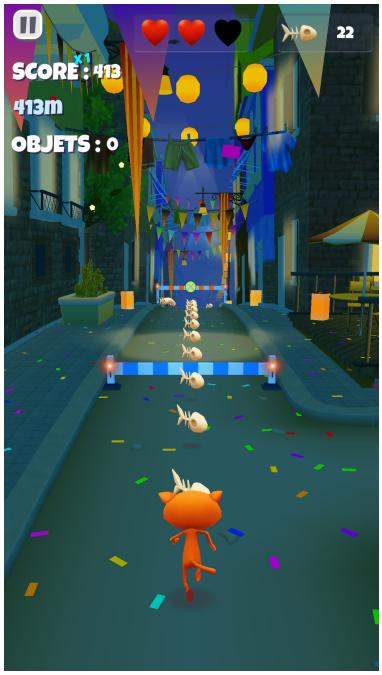


Figure 1. A screenshot of the Infinite Runner used for the experiment.

The game exists in two versions: one with binaural sounds and one in mono. Half of the sessions is presented to the subjects with a binaural rendering, the other half with a mono rendering (see Table 1). Relying on the results from a previous study [22], we set the point of listening of the binaural version to the position of the virtual camera. The sound source positions are set by using a unique set of non-individual HRTF measured from a dummy head Neumann KU 100.

Several areas have been developped: forest (spring and autumn), city (standard and snowy) and seaside town (standard and carnival), all available in a day and a night ver-

doi:10.25836/sasp.2019.23

sion. As their purpose is only to bring some graphical and sound variety, they are not subject to a precise distribution across sessions. At each session, one of them is picked up randomly and presented to the subject.

3.4 Measuring immersion, memorization and performance

Immersion, memorization and performance are measured via different methods. Immersion is measured via a questionary at the end of the session. Still keeping the idea of a short session, we elaborate four questions, inspired from previous studies on the feeling of immersion in virtual reality [23] or with spatialized sound systems [24]: 1) In the game generated world, you had a sense of "being there", 2) Have you experienced a "3D sound" effect?, 3) Sound did contribute to your feeling of immersion and 4) The external context (ambiant noise, parallel task, etc.) troubled your immersion. Except for the question 2 with a yes/no answer, all responses are given on a discret scale between 0 and 3, 0 being tagged as "not at all", and 3 being tagged as "fully agree".

Memorization is measured via a task: during the game, various audiovisual objects are randomly spread along the roadside. The objects are randomly picked up among a list of twenty, and are designed to be graphically and aurally emphasised in the scenery, to ease their memorization. At the end of the session, the last seven objects that have been met during the game are presented in a random order to the subjects, and they are asked to sort them.

Immersion and memorization both require an answer from the subjects at the end of the session. In order to prevent the subject to anticipate them (bringing a potential bias in their responses), over the 70 sessions, 30 sessions have the immersion questionary only, 30 others have the memorization task only, and the 10 remaining ones don't have any questionary (see Table 1).

Finally, performance is computed as a mix between the bonus collected and the distance traveled in the game.

In this experiment we favor external validity rather than internal validity, and we expect this choice to bring noise in our experimental data. By measuring the contribution of binaural with three different methods (a subjective questionary for immersion, a task for memorization and interactivity data for performance), we aim to obtain different kind of information, one of the method being hopefully less impacted by the influence factors.

3.5 Summary: distribution of the sessions

Table 1 summarizes the distribution of the 70 sessions considering the type of audio rendering, the period of the day and the end session questionary. In the experiment, all the sessions were presented in a random order to each subject.

Number of sessions	Sound type	Half day	End questionary
7	binaural	morning	immersion
8	binaural	afternoon	immersion
8	binaural	morning	memorization
7	binaural	afternoon	memorization
2	binaural	morning	nothing
3	binaural	afternoon	nothing
8	mono	morning	immersion
7	mono	afternoon	immersion
7	mono	morning	memorization
8	mono	afternoon	memorization
3	mono	morning	nothing
2	mono	afternoon	nothing

Table 1. Distribution of the 70 ESM sessions.

3.6 Subjects

Thirty subjects (8 women) take part in the experiment, aged between 16 and 67 (average 28.1). Among them, five are unpaid volonteers from the laboratory, the other twenty-five are externals and receive a 50 EUR voucher at the end of the experiment, to compensate for their time. The experiment are conducted in compliance with the Declaration of Helsinki as well as national and institutional guidelines for experiments with human subjects.

Subjects are recruited based on the following requirements: owning headphones and a smartphone that runs on Android OS, and having a minimal knowledge about mobile video games. Before starting the core experiment, they all are invited to come at the laboratory to get the game installed on their phone (except for five subjects who install the game by themselves, following instructions via videoconferencing). Subjects are also informed about the details of the experiment, but not about its purpose. They are instructed to accomplish their sessions as much as possible immediately after receiving the SMS notification. However, to make the experiment more fluid, any delay or anticipation are permitted within the half day of the session. Exceeding the half day, the session is postponed to the next day, extending the whole experiment at the same time.

3.7 Hypotheses

Comparing mono to binaural sessions, we provide the following hypotheses:

- for immersion, we expect better responses in the binaural sessions, resulting in a better sense of presence, a sound effect more often experienced, a better contribution of sound to the feeling of immersion and an external context less disturbing;
- for memorization, we expect objects to be better memorized in the binaural sessions;
- for performance, we expect a better global score in the binaural sessions.

4. RESULTS

Results comprise 2100 sessions (70 sessions for 30 subjects). For technical reasons, twelve sessions have been

lost, owned by a unique subject, resulting in 2088 sessions. Among them, 896 are provided with an immersion questionary, 894 with a memorization questionary and 298 sessions with no questionary.

4.1 Immersion

Figure 2 shows the answer of the subjects to the four questions related to their feeling of immersion. We observe that on average, subjects experienced a better sense of presence (question 1), a better contribution of sound to immersion (question 3) and a lower trouble caused by the external context (question 4) when audio was rendered in binaural. The answers to the question 2 reveal that most of the time, the subjects detected a 3D sound effect, whatever the audio type is, binaural or mono (366 detections of a 3D sound effect in binaural sessions, i.e., 83% of all the binaural sessions, and 349 detections in mono sessions, i.e., 76% of all the mono sessions).

Three ANOVA were performed, with the answers of questions 1, 3 and 4 as successive dependent variables, the audio rendering type as a within-subject factor, and the subject as a random factor. Results indicate a significant effect of the audio type on the sense of presence ($F(1, 29.03)=6.31$, $p<0.05$), on the contribution of sound to immersion ($F(1, 29.04)=5.05$, $p<0.05$), but not on the trouble caused by the external context ($F(1, 29.06)=1.09$, $p=0.30$). Finally, a χ^2 test is performed for question 2, with the answer type (yes or no) as the dependent variable and audio rendering type as the independent variable. Results reveal a highly probable reject of the null hypothesis ($\chi^2=6.81$, $p<0.01$), meaning that the audio rendering type has a statistically significant effect on the answers. In other words, subjects detect significantly more often a 3D sound effect when sessions are in binaural.

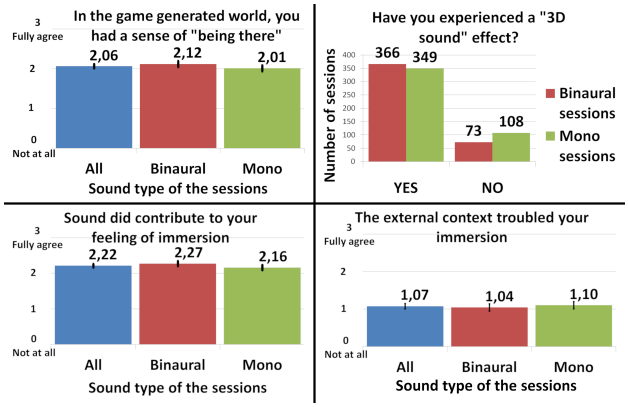


Figure 2. Subjects' answers to the four immersion related questions. The answers are systematically given in function of the type of audio rendering (mono or binaural). The answers were given on a discrete scale between 0 and 3, except for the question 2 (on the top right) where the answer was yes or no. Vertical bars are the 95% confidence intervals.

4.2 Memorization

For memorization, we have to compare the sequence of objects memorized by the subjects with the correctly ordered sequence. We compare them following several metrics: the Hamming distance [26], the Levenshtein distance [27], the Damerau-Levenshtein distance [28], the longest common subsequence [29], and the Jaro similarity [30]. The basic principle of these metrics is to compare what objects are in common in the two sequences (their number, their distance), and how many operations are required to convert one sequence to another. More details can be found in the papers associated to each method. Figure 3 shows the average normalized distances, given the audio type of the sessions. All data represent here a distance, i.e., a value of 0 if sequences are the same (good memorization), a value of 1 if sequences are very different (bad memorization). We observe for each distance that values are nearly the same for both audio renderings, but systematically lower for the mono sessions. Five ANOVA are performed, with the distances as successive dependent variables and the audio rendering type as a within-subject factor and the subject as a random factor. No significant effect of the audio rendering type is revealed, whatever the distance is: Hamming ($F(1, 29.24)=0.10$, $p=0.76$), Levenshtein ($F(1, 29.22)=0.71$, $p=0.41$), Damerau-Levenhstein ($F(1, 29.21)=0.45$, $p=0.51$), longest common subsequence ($F(1, 29.28)=2.89$, $p=0.10$) and Jaro distance ($F(1, 29.21)=0.34$, $p=0.56$).

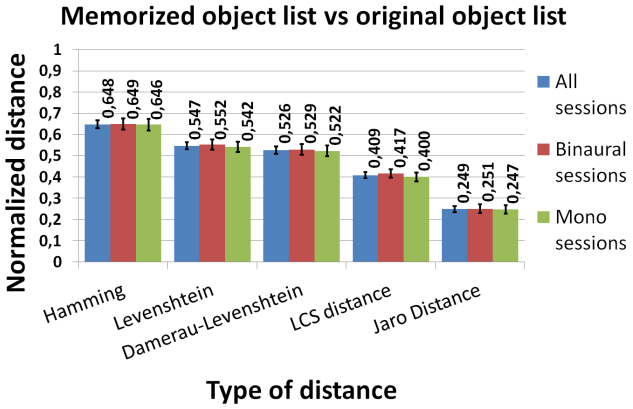


Figure 3. Average normalized distances between the memorized sequences of objects and the correctly sorted sequences. A distance of 0 means that the two sequences are identical (good memorization), and a distance of 1 means that the two sequences are highly different (bad memorization). Vertical bars are the 95% confidence intervals.

4.3 Performance

Performance of the subjects as a function of the audio rendering type can be observed in Figure 4. Values are better for binaural sessions, but here again, an ANOVA performed on data reveals no effect of the audio type ($F(1, 28.05)=0.18$, $p=0.68$).

doi:10.25836/sasp.2019.23

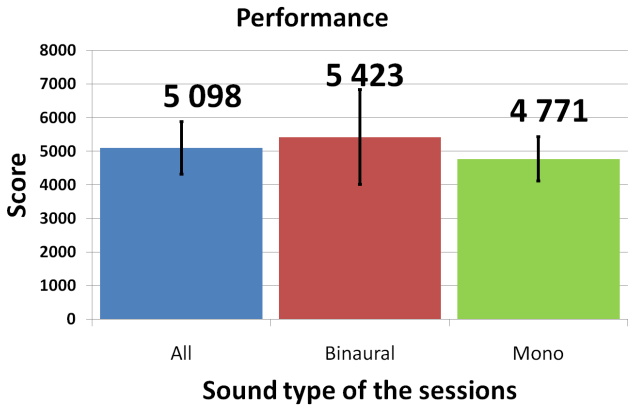


Figure 4. Vertical bars are the 95% confidence intervals.

5. DISCUSSION

Among all the questions and variables, it seems that only the sense of presence and the contribution of sound to immersion are significantly improved by the binaural rendering. Besides, a 3D effect is more often experienced in binaural sessions, meaning that subjects consciously perceive the sound spatialization.

On the contrary, memorization is not influenced by the audio rendering. One possible reason is the difficulty of the task: remembering the order of 7 objects is probably too hard, considering that 7 is usually seen as the upper limit of the memory span [31]. This was confirmed by several subjects after the experiment, during an informal debriefing. We also raise the idea that we might have mitigated the beneficial effect of the binaural rendering, by positioning objects only on the right or left of the avatar's track, instead of all around the user.

For the performance, no effect of the audio rendering type was neither found. However, many influence factors come under consideration in this experiment. First, we suppose that the difficulty of the game (and consequently performance) is probably related to the context of use, which was highly variable and not possible to study systematically in our experiment. Second, several subjects confessed having stopped their sessions intentionally, for three different reasons: some of them found that the game was too easy; some others had sometimes to stop for external reasons (end of the bus ride, end of the break at work, etc.); and some others, focused on the memorization task, died purposely to keep in mind the 7 memorization objects. In these conditions, the interpretation of the performance regarding the binaural rendering is quite difficult. In further studies, the conflict between the memorization task and the game itself may be softened by integrating the former in the core gameplay of the latter.

Finally, among the three assessed attributes though, having a significant improvement of immersion in the binaural condition is an important result. It means that in realistic use cases, despite the large variability of contexts (different places, time, situations, smartphones, headphones, inter-individual differences, etc.), binaural rendering still has a positive effect, justifying its implementation in a smartphone application.

6. CONCLUSION

We proposed an experiment to measure the contribution of a binaural rendering in a smartphone video game application. The video game was an Infinite Runner, largely spread among the general public. We measured three attributes, i.e., immersion, memorization and performance, using the Experience Sampling Method, that aims to replace the experiment in realistic contexts of use. Results revealed a better feeling of immersion with the binaural rendering, compared to the same game in mono. No significant effect was found for memorization and performance. As the study focused on external validity (realistic conditions) rather than internal validity (controlled conditions), we claim that this result justifies the use of binaural sounds in this kind of smartphone applications.

Many prospects are possible. Considering the contribution of the binaural sound, further studies should assess other attributes, other smartphone applications, and compare binaural rendering to other sound systems (e.g., stereo). Other studies may focus on the method itself, investigating how to gradually control more conditions, while keeping the external validity at the same thime.

7. REFERENCES

[1] J. Blauert, *The technology of binaural listening*. Springer, 2013.

[2] F. L. Wightman and D. J. Kistler, "Headphone simulation of free-field listening. ii: Psychophysical validation," *The Journal of the Acoustical Society of America*, vol. 85, no. 2, pp. 868–878, 1989.

[3] H. Møller, C. B. Jensen, D. Hammershøi, and M. F. Sørensen, "Using a typical human subject for binaural recording," in *Audio Engineering Society Convention 100*, Audio Engineering Society, 1996.

[4] E. H. Langendijk and A. W. Bronkhorst, "Fidelity of three-dimensional-sound reproduction using a virtual auditory display," *The Journal of the Acoustical Society of America*, vol. 107, no. 1, pp. 528–537, 2000.

[5] R. L. Martin, K. I. McAnally, and M. A. Senova, "Free-field equivalent localization of virtual audio," *Journal of the Audio Engineering Society*, vol. 49, no. 1/2, pp. 14–22, 2001.

[6] P. Larsson, D. Vastfjall, and M. Kleiner, "Better presence and performance in virtual environments by improved binaural sound rendering," in *Audio Engineering Society Conference: 22nd International Conference: Virtual, Synthetic, and Entertainment Audio*, Audio Engineering Society, 2002.

[7] A. Gonot, M. Emerit, and N. Château, "Usability of 3d-sound for navigation in a constrained virtual environment," in *AES, 120th Audio Engineering Society Convention, Paris*, Audio Engineering Society, 2006.

[8] S. Moulin, R. Nicol, and L. Gros, "Spatial audio quality in regard to 3d video," in *Acoustics 2012*, 2012.

[9] M. Cobos, J. J. Lopez, J. M. Navarro, and G. Ramos, "Subjective quality assessment of multichannel audio accompanied with video in representative broadcasting genres," *Multimedia Systems*, vol. 21, no. 4, pp. 363–379, 2015.

[10] F. Grani, F. Argelaguet, V. Gouranton, M. Badawi, R. Gaugne, S. Serafin, and A. Lecuyer, "Design and evaluation of binaural auditory rendering for caves," in *2014 IEEE Virtual Reality (VR)*, pp. 73–74, IEEE, 2014.

[11] U. Reiter, K. Brunnström, K. De Moor, M.-C. Larabi, M. Pereira, A. Pinheiro, J. You, and A. Zgank, "Factors influencing quality of experience," in *Quality of Experience*, pp. 55–72, Springer, 2014.

[12] D. R. Begault, E. M. Wenzel, and M. R. Anderson, "Direct comparison of the impact of head tracking, reverberation, and individualized head-related transfer functions on the spatial perception of a virtual speech source," *Journal of the Audio Engineering Society*, vol. 49, no. 10, pp. 904–916, 2001.

[13] C. Guastavino, *Étude sémantique et acoustique de la perception des basses fréquences dans l'environnemet sonore urbain*. PhD thesis, Université Paris 6, 2003.

[14] T. Walton, M. Evans, D. Kirk, and F. Melchior, "Exploring object-based content adaptation for mobile audio," *Personal and Ubiquitous Computing*, vol. 22, no. 4, pp. 707–720, 2018.

[15] S. Werner and F. Klein, "Influence of context dependent quality parameters on the perception of externalization and direction of an auditory event," in *Audio Engineering Society Conference: 55th International Conference: Spatial Audio*, Audio Engineering Society, 2014.

[16] A. Fiebig, "Influence of context effects on sound quality assessments," in *Proceedings of EuroNoise*, pp. 2555–2560, 2015.

[17] T. Walton and M. Evans, "The role of human influence factors on overall listening experience," *Quality and User Experience*, vol. 3, no. 1, p. 1, 2018.

[18] F. Scriven, "Two types of sensory panels or are there more?," *Journal of sensory studies*, vol. 20, no. 6, pp. 526–538, 2005.

[19] R. Larson and M. Csikszentmihalyi, "The experience sampling method," in *Flow and the foundations of positive psychology*, pp. 21–34, Springer, 2014.

[20] S. Ickin, K. Wac, M. Fiedler, L. Janowski, J.-H. Hong, and A. K. Dey, "Factors influencing quality of experience of commonly used mobile applications," *IEEE Communications Magazine*, vol. 50, no. 4, pp. 48–56, 2012.

[21] J. Traer and J. H. McDermott, "Statistics of natural reverberation enable perceptual separation of sound and space," *Proceedings of the National Academy of Sciences*, vol. 113, no. 48, pp. E7856–E7865, 2016.

[22] J. Moreira, L. Gros, R. Nicol, and I. Viaud-Delmon, "Spatial auditory-visual integration: The case of binaural sound on a smartphone," in *Audio Engineering Society Convention 145*, Audio Engineering Society, 2018.

[23] B. G. Witmer and M. J. Singer, "Measuring presence in virtual environments: A presence questionnaire," *Presence*, vol. 7, no. 3, pp. 225–240, 1998.

[24] R. Nicol, L. Gros, C. Colomes, E. Roncière, and J. Messonier, "Etude comparative du rendu de différentes techniques de prise de son spatialisée après binauralisation," *CFA/VISHNO*, 2016.

[25] N. Liu, Y. Liu, and X. Wang, "Data logging plus e-diary: towards an online evaluation approach of mobile service field trial," in *Proceedings of the 12th international conference on Human computer interaction with mobile devices and services*, pp. 287–290, ACM, 2010.

[26] R. W. Hamming, "Error detecting and error correcting codes," *The Bell system technical journal*, vol. 29, no. 2, pp. 147–160, 1950.

[27] V. I. Levenshtein, "Binary codes capable of correcting deletions, insertions, and reversals," in *Soviet physics doklady*, vol. 10, pp. 707–710, 1966.

[28] F. J. Damerau, "A technique for computer detection and correction of spelling errors," *Communications of the ACM*, vol. 7, no. 3, pp. 171–176, 1964.

[29] L. Bergroth, H. Hakonen, and T. Raita, "A survey of longest common subsequence algorithms," in *Proceedings Seventh International Symposium on String Processing and Information Retrieval. SPIRE 2000*, pp. 39–48, IEEE, 2000.

[30] M. A. Jaro, "Advances in record-linkage methodology as applied to matching the 1985 census of tampa, florida," *Journal of the American Statistical Association*, vol. 84, no. 406, pp. 414–420, 1989.

[31] G. A. Miller, "The magical number seven, plus or minus two: Some limits on our capacity for processing information.," *Psychological review*, vol. 63, no. 2, p. 81, 1956.

doi:10.25836/sasp.2019.23

DISCRIMINATION EXPERIMENT OF SOUND DISTANCE PERCEPTION FOR A REAL SOURCE IN NEAR-FIELD

Zhenyu GUO[1] **Yigang LU**[1] **Liliang WANG**[1]
Guangzheng YU[1]
[1] School of Physics and Optoelectronics, South China University of Technology, Guangzhou, China
correspondent: scgzyu@scut.edu.cn

ABSTRACT

The ability of distance perception is quite important for our daily life. It helps listeners to perceive an approaching sound source and avoid dangerous objects especially when the vision is unavailable.
Previous researches have proved that the sound pressure has a giant influence on the ability of distance discrimination in both the near field and the far field. However, a few researches attempt to examine the binaural effect alone in distance perception.
To verify the impact of binaural effect on distance discrimination, we conducted an experiment to exam the sound distance perception thresholds via an automatic test system. A loudness-balanced wide band noise was used as test signals to remove the influence of sound level. 5 azimuths (0°, 45°, 90°, 135° and 180°) and 2 reference distance (50 cm and 100 cm) are taken into consideration.
The results show that distance discrimination thresholds of subjects are lower when the sound source is on the side of head compared with front and back. Moreover, this phenomenon is more prominent in 50 cm compared with 100 cm. The results obtained in this study are consistent with previous studies and reveal that the binaural effect indeed contributes to distance discrimination process of human to some degree.

1. INTRODUCTION

Spatial hearing is a vital ability for human beings to percept objects especially when vision system is disabled. Not only spatial angle but also distance of sound sources can be perceived by people when referring to spatial sound location. Vast researches have studied the spatial angle perception of humans [1], however, less literature focus on inquiring distance discrimination ability of humans.
According to previous researches, the following several acoustic cues are considered to be the most vital for distance perception: (1) intensity; (2) direct-to-reverberation energy ratio; (3) spectrum change; (4) dynamic cues; (5) binaural cues [2, 3]. Among all these cues, intensity is the most significant and general one especially in the far field [4]. Since the shadow effect of head could be more significant in the near field, we believe the binaural cues can also be a useful cue for sound source distance discrimination. The propagation process from sound source to two ears can be described by the head-related transfer function (HRTF). Normally, HRTFs contain interaural level difference (ILD), interaural time difference (ITD) and spectrum cues [5]. Both theoretical calculation and measurement demonstrate that distinction of HRTFs between two ears change quite significantly with distance in the near field [6, 7]. For instance, the ratio of two ears' ILD changes with distance and azimuth of sound source dramatically especially in the near field. This is because that shadow effect of head contributes a great deal in the near field. On one hand, the head can be regarded as a rigid reflection panel, and it attenuates the sound wave propagating to the contralateral ear, on the other hand, it will enhance sound wave propagating to the ipsilateral ear due to reflection especially for high frequency components of sound wave. As we know, high-frequency sound is highly directional, in this case high-frequency sound's mask degree can change with sound source distance. Besides, the azimuth related to the lateral ear of sound source vary with its distance to head center, and this is called acoustic parallax [8]. Considering above-mentioned facts, we have enough reasons to believe that distance discrimination ability could make a difference when sound source appears in different lateral azimuths.
Up to now, there were a few literatures conducting experiments to verify the head shadow effect in distance discrimination, or in other words, effect of binaural cue in distance discrimination. Brungart et al. have examined the distance discrimination performance of lateral sound source [9]. However, their experiment failed to exclude other distance perception cues. Therefore, it is necessary to carry out a specific and systematic experiment to examine to what extent binaural cues can make effect in sound source distance perception.
Generally speaking, two categories of sound source are alternative to conduct psychoacoustic experiment. One is virtue auditory display (VAD) sound source [10], while the other is real sound source [9]. In terms of VAD via headsets, signal must be convolved with individual HRTFs with the intention of achieving real spatial perception. Nev-

© Zhenyu GUO, YiGang LU, Liliang WANG, Guangzheng YU. Licensed under a Creative Commons Attribution 4.0 International License (CC BY 4.0). **Attribution:** Zhenyu GUO, YiGang LU, Liliang WANG, Guangzheng YU. "Discrimination experiment of sound distance perception for a real source in near-field", 1st EAA Spatial Audio Signal Processing Symposium, Paris, France, 2019.

doi:10.25836/sasp.2019.25

ertheless, the accuracy of near-field HRTFs measurement still remains problems [11]. Besides, the coupling problem between headsets and ear canals can be quite tough and uncontrollable, and this problem may influence binaural ITDs a lot. In order to assess the distance discrimination in an environment as real as we can, we choose to use a loudspeaker to playback test signal. We built a moveable loudspeaker playback system by using an electric slideway for the experiment and conducted the complete experiment in an anechoic chamber (background noise below -12.1 dBA) to exclude the impact of reverberation or any other possible hint cues.

2. METHODS

The distance discrimination experiment was conducted with a specially designed platform in an anechoic chamber. The subjects' relative discrimination thresholds in five different azimuths (0°, 45°, 90°, 135°, 180°, in the right half plane) and two distinct reference distances (d_f, 50 cm or 100 cm) were measured via two-interval forced choice (2IFC) [12]. The discrimination performance was evaluated in two distinct conditions (with the intensity cue excluded or included, considered as the experimental group and the control group, respectively). To make a summary, each subject's distance discrimination thresholds were measured in total 20 ($5 \times 2 \times 2$) different conditions.

2.1 Subjects

Eight subjects (four males and four females, denoted as S1 to S8) participated in the discrimination experiment. Subjects' ages range from 23 to 41. Each subject has normal hearing and has experience in participating psychoacoustic experiments. Subjects were paid for their participation.

2.2 Experiment devices

Pervious distance discrimination experiments were commonly implemented with a simple mechanical movable loudspeaker system [9, 13]. The loudspeaker is pushed by the assistant manually. Considering that the assistant may make mistakes in moving the loudspeaker and the action may create noise which may hint the subjects about sound source distance, neither the accuracy nor the reliability of these experiments' result data is doubtful. With the intention of conducting distance discrimination experiment, sound source must appear in different distances to subject rapidly and accurately. To accomplish this aim, we built an automatic moveable loudspeaker playback system. The loudspeaker (Mission M30i) is attached to an electric slide way which is driven by a two-phase stepper motor (57BYGH75). The mechanical system is fixed on a stable aluminum bracket. As for the control system, we use a commonly used embedded control board named Arduino (Arduino Uno) and a two-phase stepper motor driver (DM542). In terms of audio stream, we use an external sound card (RME Fireface UC) to output stimulus signals, an amplifier (ARCAM A65) cascades to the sound card

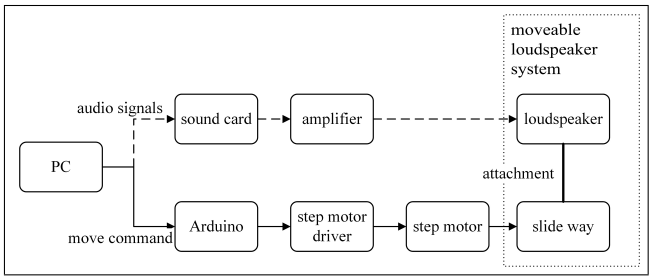


Figure 1. The schematic diagram of experiment device

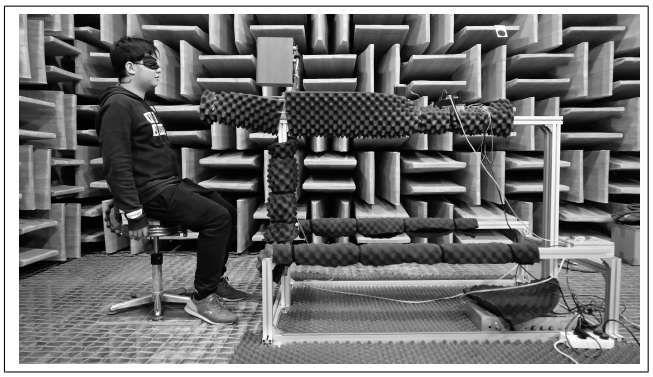


Figure 2. The picture of experiment device.

and drives the loudspeaker. The whole control flow is as follows:

1. The computer transmits move orders to Arduino through the Universal Serial Bus (USB) port. Meanwhile, the computer transmits the test signal to the audio card.
2. The Arduino receives the move order and translates it into pulse width modulation (PWM) signals and direction control driver signals for DM542.
3. DM542 transforms the PWM signals and direction control signal into two-phase stepper motor code driver signal, then the motor can put loudspeaker in motion.

The schematic diagram and picture of the device are shown in Figure 1, and Figure 2, respectively. The valid move range of loudspeaker can reach to 0.8 m, maximum move speed is 500 mm/s, and the error of move distance is no more than 0.1mm. The whole system is wrapped with sound-absorbing cotton to avoid unnecessary reflection. Although the slide way will make a little noise (less than 35 dBA) during operating, it cannot provide subjects the distance and movement statue information of loudspeaker according to noise evaluation and subjects' feedback.

2.3 Stimuli and self-adaptive procedure

All stimuli used in this study are full-band pink noise. Signal sampling rate is 44.1 kHz, duration time is 1000 ms and the ramp time is 20 ms. Two test conditions are set to exclude the influence of intensity cue. To be specific, signals of the experimental group is balanced sound pressure

doi:10.25836/sasp.2019.25

level (SPL) while no compensation is done for signals of the control group. In terms of experimental group, SPL of stimulus in the position of head center is constant 75 dBA no matter how far the sound source is. As for the control group, the SPL of stimuli in the above-mentioned position is calibrated as 75 dBA only when sound source in reference location, and no intensity compensations are made, that means the sound pressure of sound source in the head position obeys the inverse-square law [6].

The distance discrimination thresholds of different conditions were measured via implementing 2IFC and 2-down-1-up self-adaption method. To be specific, loudspeaker will playback two stimulus in a round. Before each round, control computer plays a prompt message to hint subject. One stimulus appears in reference location while other one appears in a forward location (test distance, d_t) which is closer to the subject, besides, the order of two stimuli was random. After two stimuli are played over, subjects are required to choose which stimulus is more near to them and feedback their choices to the assistant in oral. The interval time between two stimuli ranges from 1 s to 2 s (depending on distance difference between two stimuli). In the first round of each block, the initial d_t is 30% less than the d_f when sound source's azimuth was 0°, 45°, 90° or 135°, and this value was adjusted to 40% for 180° particularly (e.g. d_t is 0.35 m for the condition that d_f is 50 cm and sound source azimuth is 0°). After subjects feedbacked their judgement, d_t would be adjusted automatically with the 2-down-1-up method. That means only when subjects fetch correct answer twice continuously will d_t be shorten a step length, on the other hand, d_t will be extended following each error responding, otherwise, d_t keeps constant in the next round. This procedure guarantees that the correction rate of subjects will not be under 70.7% [14]. With regard to the step length for each adjustment of d_t, it is 3% before the first mistake just for accelerating the convergence speed. After a mistake choice is made, the step value will become 1% (that is 0.5 cm when d_f is 50 cm or 1cm when d_f is 100 cm) for subsequent test rounds. Particularly, if subjects fetch correct responding four times in a row, the step length is adjusted to 2% until next mistake appears. The whole self-adaptive procedure finish when action of decreasing or increasing d_t reverses 12 times. The final discrimination distance d_t' is determined on the mean value of d_t in the last 5 reversal rounds.

The distance discrimination threshold was measured by Weber ratios [15] as following:

$$threshold = \frac{d_f - d_t'}{d_f} \tag{1}$$

threshold under 20 discrimination condition of each subject were measured. Figure 3 presents a round of discrimination threshold measurement.

2.4 Procedure

The experiment procedure for each subject is depicted as followings:

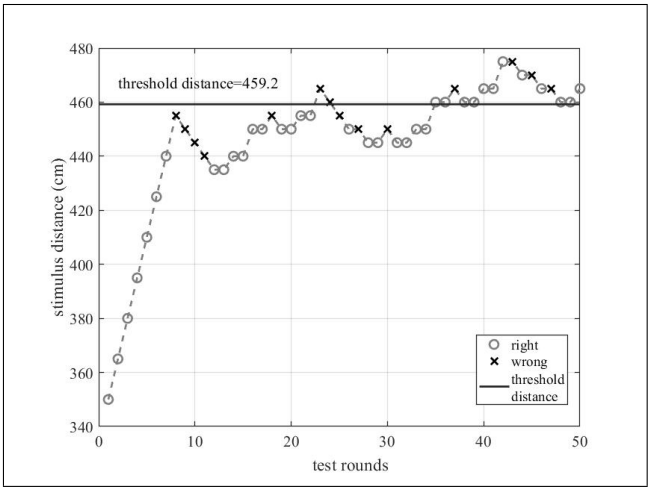


Figure 3. A block of discrimination threshold measurement. In this test, d_t' is 459.2 cm while *threshold* is 8.2% in the condition that reference distance is 500 cm and the intensity of stimulus is unbalanced.

1. Choose one from five test azimuth in random, and adjust the subject's azimuth relative to slideway. After that, we asked the subject to wear a blinder to guarantee they would not get any clues from vision.
2. Select one reference distance between two alternative choices.
3. Measure the *threshold* values of control group and experimental group in sequence by the scheme described in section 2.3.
4. Repeat above-mentioned process until *threshold* in total 20 different conditions was measured completely..

The full test for each subject consumed about 6 hours, and each subject only participated the experiment for 2 hours a day, besides, subjects had 10 minutes for rest after completing each test block. Before the experiment, subjects got a simple training (play ten rounds of stimuli) to familiarize themselves with the stimuli.

3. RESULTS AND DISCUSSIONS

3.1 Results of *threshold* measurement

8 participators' *threshold* in all condition was obtained in this experiment and presented in Figure 4, in this figure, black line and gray line indicates the control group and experimental group, respectively. Besides, the circle marks and the square marks represent the 50 cm and 100 cm reference distance, respectively.

3.2 Intensity cue

There is no doubt that the intensity cue is the primary cue for sound source distance prescription. The results verify this conclusion again considering that *threshold* values in unbalanced intensity condition are 10% to 20%

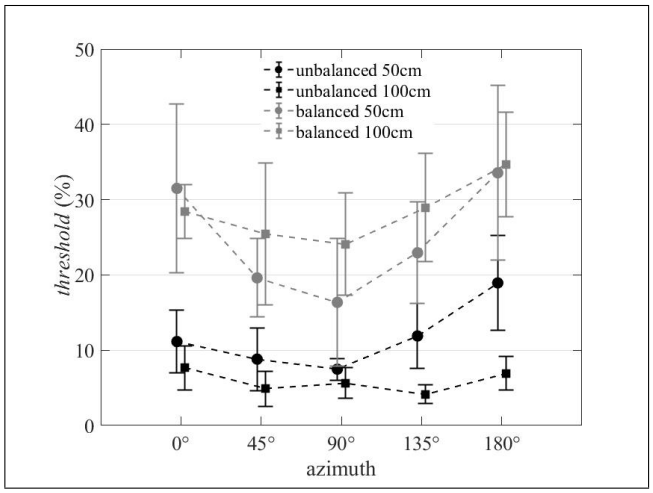


Figure 4. *threshold* measurement results. The marks and the caps indicate the mean values and the standard deviation values among 8 subjects, respectively.

lower than values in balanced intensity condition. Conducting multivariate analysis of variance (ANOVA) which include three factors as mentioned in section 2, result comes out that intensity cue has significance in main effect ($F(1,140)=287.542$, $p<0.05$).

3.3 Reference distance

Both in experimental group and control group, reference distances make significance effect to *threshold* value. The ANOVA results come out that $F(1,140)=15.224$, $p<0.05$ and $F(1,140)=5.653$, $p=0.019$ for experimental group and control group, respectively. Especially for control group which includes intensity cue, the result demonstrates that the father the sound source is, the lower the *threshold* is. Strybel et.al. [15] and Simpson et al. [16] reported similar phenomenon before. In terms of experimental group which exclude intensity cue, the results don't meet abovementioned regulation. To be specific, the *threshold* value becomes lower rather than higher in 50 cm reference distance when sound source is lateral. This phenome reflects head shadow effect in sound source distance prescription considering that head shadow makes significance effect in ILD especially for lateral sound source in the near field and the ILD cue (or say binaural cue) can be used to estimate sound source distance by auditory system especially when the intensity cue is unavailable.

3.4 Binaural cue

When we refer to binaural cues, ILD and ITD are both necessary to be considered. However, previous study has proved that ITD provide less information in distant perception [11]. Accordingly, this study focuses on discuss the influence of ILD but not ITD in distance perception.
The experiment in this study measured 5 different azimuth of sound source in the horizontal plane to uncover whether binaural cue make effect in distance discrimination. Figure 4 demonstrates that threshold values tend to be smaller when sound source is in the lateral direction (45°, 90°, 135°) compared with the medium plane (0° and 180°), and the results reveal that distance discrimination ability is better for lateral sound source than sound source in the medium plane and some researches have similar conclusion as well [9, 10].
This pattern appears both in experimental group and control group, especially for the case which excludes the intensity cue. That is due to that the ILD change with sound source distance dramatically in the near field. As the sound source becomes closer to listener, the ILD becomes larger gradually. This ILD changing pattern has been calculated via analytical solution of the rigid sphere model [6]. Considering sound source appear in medium plane, the loudness of source in two ears would be the same hence no ILD cue is available. On one hand, ILD cue which caused by head shadow effect become a main distance perception cue and lead to huge threshold distinction between lateral and front (or rear) direction in the experimental group as the gray curve in Figure 4 shows. On the other hand, ILD become a supplementary cue considering that *threshold* value increase below 5% when sound source appears in lateral direction comparing with medium plane.
As the reference distance decreases, the head shadow effect will become more significant, or in other words, the ILD cue become more obvious. This object cue also makes sense in auditory perception according to Figure 4. The round markers represent 50 cm reference distance while the square markers denote 100 cm reference distance. The curve of 50 cm changes with different azimuths more comparing with the curve of 100 cm especially when intensity cue is unavailable.

4. CONCLUSIONS

The study built a fast and accurate experiment platform for measuring distance discrimination threshold of human beings. The intention of this study is to reveal to what degree can the binaural cue make impact on distance discrimination. 5 different azimuths from 0° to 180° with a step of 45°, 2 discrete reference distances and 2 types of stimuli are taken into considering. Utilizing the experiment platform, total 160 *threshold* values (8 subjects, 20 *threshold* values in different conditions for each subjects) are measured.
The results demonstrate intensity cue is indeed the main cue for distance discrimination. Meanwhile, the reference distance also makes effect in distance discrimination, that is, particularly speaking, the father the sound source is, the lower the distance discrimination threshold is when sounds include intensity cue.
In addition, the results reveal that the binaural cues definitely have influence on sound source distance perception:

1. The lateral direction distance discrimination ability is better compared with medium plane

2. The binaural cue is less important when intensity cue is available although it indeed makes influence.

doi:10.25836/sasp.2019.25

3. The influence of binaural cue on distance discrimination increases when sound source gets closer.

5. ACKNOWLEDGMENTS

The research is supported by National Natural Science Foundation of China (Grant No. 11574090) and the Natural Science Foundation of Guangdong Province (Grant No. 2018B030311025). The anechoic chamber and audio devices are provided by School of Architecture and School of Physics and Optoelectronics, South China University of Technology, respectively.

6. REFERENCES

[1] M. J. H. Agterberg, A. F. M. Snik, M. K. S. Hol, M. M. Van Wanrooij, and A. J. Van Opstal, "Contribution of monaural and binaural cues to sound localization in listeners with acquired unilateral conductive hearing loss: Improved directional hearing with a bone-conduction device," *Hearing Research*, vol. 286, no. 1-2, pp. 9–18, 2012.

[2] D. H. Ashmead, D. Leroy, and R. D. Odom, "Perception of the relative distances of nearby sound sources," *Perception & psychophysics*, vol. 47, no. 4, p. 326, 1990.

[3] X. Bosun, *Head-related transfer function and virtual auditory display*. USA: J.Ross Publishing, 2013.

[4] D. S. Brungart, "Auditory parallax effects in the hrtf for nearby sources," 1999.

[5] D. S. Brungart, N. I. Durlach, and W. M. Rabinowitz, "Auditory localization of nearby sources. ii. localization of a broadband source," *The Journal of the Acoustical Society of America*, vol. 56, no. 6, pp. 1829–1834, 1974.

[6] D. S. Brungart and W. M. Rabinowitz, "Auditory localization of nearby sources. head-related transfer functions," *The Journal of the Acoustical Society of America*, vol. 106, no. 3, pp. 1465–1479, 1999.

[7] P. D. COLEMAN, "An analysis of cues to auditory depth perception in free space," *Psychol Bull*, vol. 60, pp. 302–15, 1963.

[8] A. J. Kolarik, B. C. J. Moore, P. Zahorik, S. Cirstea, and S. Pardhan, "Auditory distance perception in humans: a review of cues, development, neuronal bases, and effects of sensory loss," *Attention, Perception, & Psychophysics*, vol. 78, no. 2, pp. 373–395, 2016.

[9] N. Kopčo and B. G. Shinn-Cunningham, "Effect of stimulus spectrum on distance perception for nearby sources," *The Journal of the Acoustical Society of America*, vol. 130, no. 3, pp. 1530–1541, 2011.

[10] H. Levitt, "Transformed up-down methods in psychoacoustics," *The Journal of the Acoustical Society of America*, vol. 49, no. 2B, pp. 467–477, 1971.

[11] M. Otani, T. Hirahara, and S. Ise, "Numerical study on source-distance dependency of head-related transfer functions," *The Journal of the Acoustical Society of America*, vol. 91, no. 3, pp. 1637–1647, 2009.

[12] W. E. Simpson and L. D. Stanton, "Head movement does not facilitate perception of the distance of a source of sound," *The American Journal of Psychology*, vol. 1, no. 86, pp. 151–9, 1973.

[13] I. Spiousas, P. E. Etchemendy, M. C. Eguia, E. R. Calcagno, E. Abregú, and R. O. Vergara, "Sound spectrum influences auditory distance perception of sound sources located in a room environment," *Frontiers in Psychology*, vol. 8, 2017.

[14] T. Z. Strybel and D. R. Perrott, "Discrimination of relative distance in the auditory modality: The success and failure of the loudness discrimination hypothesis," *The Journal of the Acoustical Society of America*, vol. 76, no. 1, pp. 318–320, 1984.

[15] N. F. Viemeister, "Temporal modulation transfer functions based upon modulation thresholds," *The Journal of the Acoustical Society of America*, vol. 66, no. 5, pp. 1364–1380, 1979.

[16] P. Zahorik, D. S. Brungart, and A. W. Bronkhorst, "Auditory distance perception in humans: A summary of past and present research," *ACTA ACUSTICA UNITED WITH ACUSTICA*, 2005.

AUDITORY VERTICAL LOCALIZATION IN THE MEDIAN PLANE WITH CONFLICTING SPECTRAL AND DYNAMIC CUES

Bosun Xie
Acoustic lab, School of Physics and Optoelectronics, South China University of Technology, Guangzhou, China
phbsxie@scut.edu.cn

Jianliang Jiang
Acoustic lab, School of Physics and Optoelectronics, South China University of Technology, Guangzhou, China
403738218@qq.com

Chengyun Zhang
School of Mechanical and Electrical Engineering, Guangzhou University, Guangzhou, China
zhangcy@gzhu.edu.cn

Lulu Liu
Acoustic lab, School of Physics and Optoelectronics, South China University of Technology, Guangzhou, China
348802659@qq.com

ABSTRACT

Both spectral and dynamic cues provide information for front-back and vertical localization. However, the relative importance of two cues to vertical localization is still unclear. The spectral cue has conventionally been regarded as a dominant one and the contribution of dynamic cue was often neglected. In present work, a psychoacoustic experiment was conducted to examine the relative importance of spectral and dynamic cues to vertical localization in the median plane. By modifying the head-related transfer functions (HRTFs) used in a dynamic virtual auditory display, binaural signals with conflicting spectral cue and dynamic variation of interaural time difference (ITD) were created and presented by headphone. Results indicated that dynamic cue dominates the vertical localization below the frequency of 3 kHz, and the spectral cue dominates the vertical localization above the frequency of 3 kHz. For stimuli with full audible bandwidth, conflicting spectral and dynamic cues results in two splitting auditory events or perceived virtual sources corresponding to different frequency band. The vertical positions of high-frequency and low-frequency virtual source are consistent with the spectral cue and dynamic cues, respectively. Therefore, both spectral and dynamic cues are important to vertical localization but each is effective at different frequency range.

1. INTRODUCTION

It has been well established that binaural cues including ITD (interaural time difference) below the frequency of 1.5 kHz and ILD (interaural level difference) at high frequency account for lateral auditory localization. It is also known that spectral cue, especially that caused by pinna at high frequency above 5 ~ 6 kHz contributes to front-back and vertical localization [1-4]. In addition, early in 1940, Wallach hypothesized that the dynamic cue caused by head turning provides information for front-back and vertical localization [5]. In detail, the ITD variation caused by head turning around the vertical axis (rotation) allows the discrimination of the front-back location as well as the determination of vertical displacement from the horizontal plane. The ITD variation caused by head turning around the front-back axis (tilting) provides supplementary information for up-down discrimination. Wallach also reported an experimental evidence for his hypothesis. After Wallach, many experiments further validated the contribution or even dominant role of dynamic cue in front-back discrimination at low frequencies [6-7]. In contrast, there have only been a few experiments (including some modern experiments) to verify Wallach's hypothesis on vertical localization [8].

Actually, auditory localization is the consequence of comprehensive processing of multiple localization cues by the high level nervous system [1-2]. In the case of an actual sound source, multiple localization cues provide consistent information and enhance the accuracy in localization. In some cases of spatial sound reproduction, however, some vertical localization cues may be absent (such as in the case of stable virtual auditory display), error, or even conflicting [9]. Therefore, it is necessary to examine the contribution of spectral and dynamic cues to vertical localization. However, the relative importance of two cues is still unclear. The spectral cue has conventionally been regarded as a dominant one and the contribution of dynamic cue was often neglected.

In present work, the ITD variation caused by head turning was first analyzed; then a psychoacoustic experiment was conducted to examine the vertical localization in the median plane with conflicting spectral and dynamic cues.

© Bosun Xie, Jianliang Jiang, Chengyun Zhang, Lulu Liu. Licensed under a Creative Commons Attribution 4.0 International License (CC BY 4.0). **Attribution:** Acoustic lab, School of Physics and Optoelectronics, South China University of Technology, Guangzhou, China; Acoustic lab, School of Physics and Optoelectronics, South China University of Technology, Guangzhou, China; School of Mechanical and Electrical Engineering, Guangzhou University, Guangzhou, China; Acoustic lab, School of Physics and Optoelectronics, South China University of Technology, Guangzhou, China;. "Auditory vertical localization in the median plane with conflicting spectral and dynamic cues", 1st EAA Spatial Audio Signal Processing Symposium, Paris, France, 2019.

doi:10.25836/sasp.2019.06

2. ANALYSIS ON THE ITD VARIATION CAUSED BY HEAD TURNING

Spatial direction relative to head center is specified by interaural polar coordinate $\Omega = (\Theta, \Phi)$. Where $-90^o \le \Theta \le 90^o$ denotes the interaural polar azimuth, that is, the angle between the directional vector of the sound source and the median plane, with $\Theta = -90^o$, 0^o, and 90^o being the left direction, median plane, and right direction, respectively. The $-90^o \le \Phi < 270^o$ denotes the interaural polar elevation, that is, the angle between the projection of the directional vector of the source to the median plane and the frontal axis, with $\Phi = -90^o$, 0^o, 90^o, and 180^o being the below, front, above, and back directions, respectively.

There are various definitions and methods for ITD calculation [2]. In present work, the ITDs are calculated by maximizing the normalized cross-correlation function between a pair of head-related transfer functions (HRTFs). Considering that ITD is an effective localization cue at low frequency, the frequency range for the ITD calculation is up to 1.5 kHz.

The HRTFs of KEMAR artificial head were used in analysis. The HRTFs were obtained by 3D-laser-scanned anatomical model and BEM-based calculation [10]. The model included the head, pinnne and neck, but excluded the torso. HRTFs at a far-field distance of 1.5 m were calculated. The sampling frequency of resultant HRTFs was 44.1 kHz, with a frequency resolution of 50 Hz. Both azimuthal and elevation resolutions of HRTFs were 1°. Due to the large error in the calculated HRTFs at low elevation when the torso is omitted, the HRTFs at low elevation ($\Phi < -30^o$ or $\Phi > 210^o$) were neglected.

Actually, ITD is approximately zero for a sound source in the median plane. If head rotates around the vertical axis, the ITD changes. The variation of ITD depends on both source elevation and azimuth of rotation. Fig.1 (a) plots the ITD variation (ΔITD) after head rotation to the left with various angle $\Delta\theta$ and as a function of source elevation Φ in the median plane.

It is observed that the ITD variation increases with the azimuth of head rotation. For a given azimuth of head rotation, the magnitude (absolute value) of ITD variation maximizes for the sound source at the directly front ($\Phi = 0^o$) and back ($\Phi = 180^o$) directions. As the sound source departs from these two directions to the high or low elevation, the magnitude of ITD variation reduces. At the above direction ($\Phi = 90^o$), the ITD is approximately invariable against head rotation. In addition, the polarity of ITD variation is opposite between the front ($-30^o \le \Phi < 90^o$) and rear ($90^o < \Phi \le 210^o$) regions. Therefore, ITD variation caused by head rotation allows the discrimination of front-back location. It also provides information for vertical displacement of sound source from the horizontal plane. This is just the Wallach's hypothesis on front-back and vertical auditory localization.

It is also observed from Fig.1(a) that the ITD variation with head rotation is approximately up-down symmetric. For example, the ITD variations for $\Phi = 30^o$ and $\Phi = -30^o$ are approximately equal. Wallach also hypothesized that head tilting provides supplementary information for up-down discrimination. Fig.1 (b) plots the ITD variation after head tilting to left with various angles $\Delta\gamma$ and as a function of source elevation Φ in the median plane. The polarity of ITD variation is opposite for sound sources above and below the horizontal plane. This is consistent with Wallach's hypothesis.

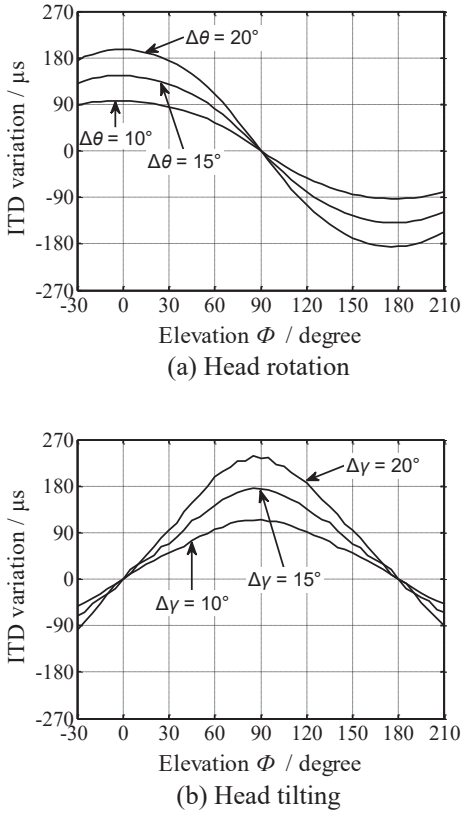


Figure 1. The ITD variation after head turning

3. EXPERIMENTAL PRINCIPLE AND METHOD

3.1 Experimental principle

For an actual sound source at a given direction in the free-field, the ITD and spectral cue for auditory localization are determined by the magnitudes and phase of binaural pressures, respectively. When the head turns, both magnitudes and phases of binaural pressures change. Accordingly, spectral cue and ITD vary consistently in terms of the new source direction relative to head.

To explore the relative contribution of dynamic ITD variation and spectral cue to localization in the median plane, inconsistent variations of ITD and the spectral cue are created. As indicated in the Section 2, for a sound source in the median plane and a given angle of head turning, the ITD variation depends on target source elevation. Therefore, the following inconsistent variations of binaural pressures should be simulated. For a target source in the median plane, the magnitudes of binaural pressures and their variation with head turning are created as if target source is located at an elevation (called spectra-based elevation Φ_{spe}); while the ITD and its variation with head turning is created as if target source is located at another elevation (called ITD variation-based elevation $\Phi_{\Delta ITD}$).

The inconsistent localization cues are created by a dynamic virtual auditory display. The binaural signals

doi:10.25836/sasp.2019.06

(pressures) of a target sound source at a target spatial direction Ω in the free field are synthesized by filtering the input stimulus E_0 with a pair of corresponding HRTFs:

$$E_\alpha = H_\alpha(\Omega, f)E_0 \quad (1)$$

Where $\alpha = L$ or R represents the left and right ears, respectively. When the head turns, the HRTFs in Eq.(1) are updated in real-time in terms of the temporary direction of target source relative to head detected by a head tracker.

Previous studies indicated that a HRTF can be approximated as its minimum-phase function cascading a linear phase (or delay) term [11]:

$$H_\alpha(\Omega, f) = H_{\min,\alpha}(\Omega, f)\exp[-j2\pi f T_\alpha(\Omega, \varphi)] \quad (2)$$

The magnitude of minimum-phase function is identical to that of original HRTF, and the phase of minimum-phase function is related to the logarithmic magnitude by Hilbert transform. The linear delays are evaluated by maximizing the normalized cross-correlation function between the minimum-phase function and original HRTF [2].

According to Eq.(2), the spectral cue in reproduced binaural pressures is determined by the minimum-phase functions of HRTFs, and the ITD is determined by the linear delay. To create inconsistent variation of ITD and the spectral cue, the HRTFs Eq.(2) are modified and then used to synthesize the binaural signals according to Eq.(1). That is, the minimum phase functions of HRTFs and their variations with head turning are stimulated in terms of spectra-based elevation Φ_{spe}; and the linear delays and their variations with head turning are stimulated in terms of ITD variation-based elevation $\Phi_{\Delta ITD}$.

3.2 Experimental method and procedures

The experiment was conducted via a PC-based dynamic VAD. An electromagnetic head tracker (Polhemus FASTRAK) detected the turning of the subject's head in three degrees of freedom and then VAD synthesized binaural signals using the method in Section 3.1. Individualized HRTFs obtained by the same method as in Section 2 were modified and then used to synthesize the binaural signals. The resultant binaural signals were reproduced by a pair of in-ear headphone (Etymotic Research ER-2). Because the ER-2 headphone exhibited a flat magnitude response measured at the end of an occluded-ear simulator, the equalization of the headphone to eardrum transmission was omitted. The update rate and system latency time of the VAD were 60 Hz and 25.4 ms, respectively. The details of the dynamic VAD are referred to [12].

Nine spectra-based elevations, ranging from Φ_{spe} = -30º to 210º at an interval of 30º in the median plane, were chosen. The ITD variation-based elevations $\Phi_{\Delta ITD}$ were chosen as follows,

(1) For each Φ_{spe} = 0º, 90º and 180º, nine $\Phi_{\Delta ITD}$, ranging from -30º to 210º at an interval of 30º in the frontal and back-median plane as well as above direction, were chosen.
(2) For Φ_{spe} = -30º, 30º and 60º in the frontal-median plane, five $\Phi_{\Delta ITD}$, ranging from -30º to 90º at an interval of 30º in the front-median plane and above direction, were chosen.
(3) For Φ_{spe} = 120º, 150º and 210º in the back-median plane, five $\Phi_{\Delta ITD}$, ranging from 90º to 210º at an interval of 30º in above direction and the back-median plane, were chosen.

Low pass-filtered noise with a cut-off frequency of 3.0 kHz and pink noise with full audible bandwidth were used as stimuli. The length of each stimulus was 5 s. The experiment was conducted in a sound-proof listening room where the level of background noise was less than 30 dBA. Binaural signals were presented at a sound pressure level equivalent to a free-field presentation of approximately 75 dBA. Eight subjects (4 male and 4 female) participated in the experiment. The subjects aged from 22 to 30 and had normal hearing.

The subjects judged the perceived virtual source direction and reported using an electromagnetic tracker (Polhemus FASTRAK) during the stimulus presentation. The tracker included two receivers. One receiver was fixed on the subject's head surface to monitor the position and orientation of the head. Another receiver was fixed at one end of a 1.0 m wooden rod. The subjects pointed the rod at the position of the perceived virtual source and a computer recorded the results. The results were transferred into the perceived direction relative to head centre. If auditory event was located inside the head (in-head localization), the subjects reported the results orally. If two splitting virtual sources or auditory events were perceived, the directions of two virtual sources were reported separately. Under each condition, each subject repeatedly judged four times. Therefore, there were 4 repetitions × 8 subjects = 32 judgments under each condition.

The mean unassigned perceived polar azimuth, mean perceived polar elevation and corresponding standard deviation are calculated across 32 judgments under each condition. Prior to calculating the mean perceived polar elevation, the judged directions for the front-back and up-down confusion cases are resolved and the percentages of confusion are calculated.

4. EXPERIMENTAL RESULTS

In all cases, no in-head-localization occurs. The mean unassigned perceived polar azimuths are less than 4º, i.e., the perceived virtual sources are located near the median plane. Therefore, we focus on the perceived elevations in the following.

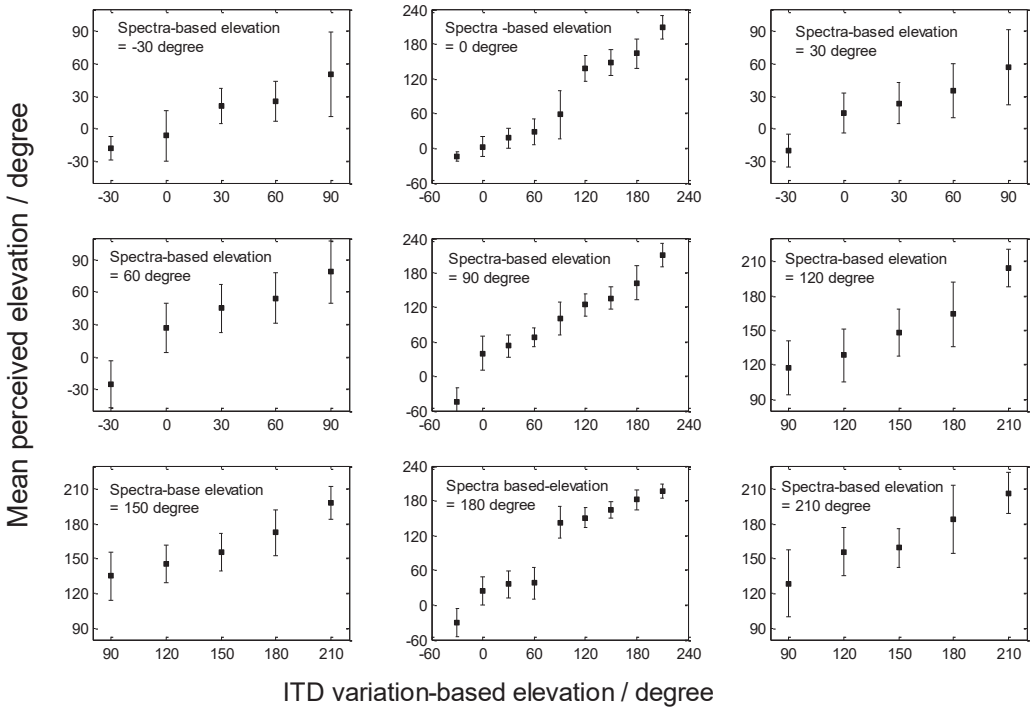


Figure 2. Mean perceived elevation for 3 kHz low-pass filtered noise

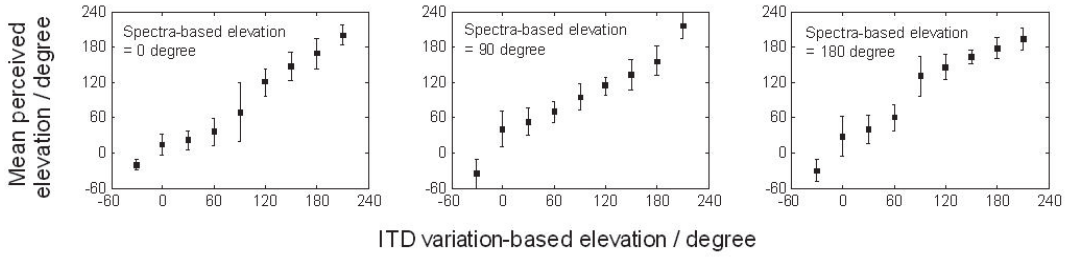


Figure 3. Mean perceived elevation for the low-frequency component of pink noise

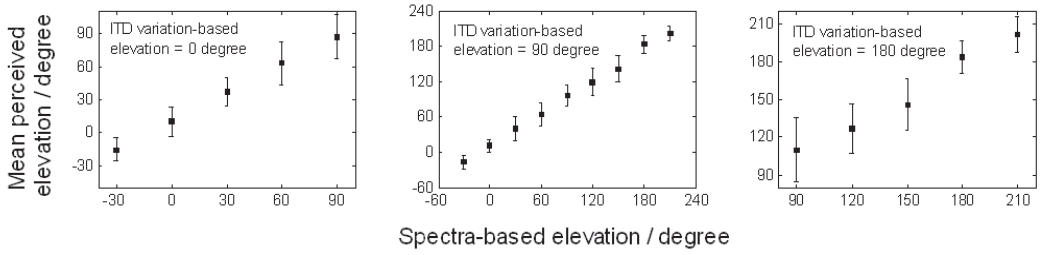


Figure 4. Mean perceived elevation for the high-frequency component of pink noise

doi:10.25836/sasp.2019.06

For the 3 kHz low-pass filtered stimulus, a single virtual source was perceived. Fig.2 shows the variation of mean perceived polar elevation with ITD variation-based elevation $\Phi_{\Delta ITD}$ for various spectra-based elevation Φ_{spe}. The results can be summarized as follows:

(1) The front-back location of perceived virtual sources basically follows that of $\Phi_{\Delta ITD}$, in spite of the Φ_{spe}. The percentages η_{FB} of front-back confusion referred to $\Phi_{\Delta ITD}$ do not exceed 10% in most cases. For examples, when Φ_{spe} = 0º (front) but $\Phi_{\Delta ITD}$ = 180º (back), the perceived elevation is near 180º (back). When Φ_{spe} = 180º (back) but $\Phi_{\Delta ITD}$ = 0º (front), the perceived elevation is near 0º (front). For a few exceptions, η_{FB} lies between 10% to 25%. Only for three cases of ($\Phi_{\Delta ITD}$, Φ_{spe}, η_{FB}) = (60º, 180º, 65.5%), (120º, 0º, 48.3%), (210º, 0º, 30%), η_{FB} exceeds a limit of 25%.

(2) The up-down locations of perceived virtual source exhibit some confusion. The percentages η_{UD} of up-down confusion referred to $\Phi_{\Delta ITD}$ vary from 0% to a high value of 78.1%.

(3) After resolving the front-back and up-down confusion, the mean perceived virtual source elevation varies basically consistent with $\Phi_{\Delta ITD}$, in spite of Φ_{spe}.

For pink noise with full audible bandwidth, two splitting virtual sources were perceived, with one corresponding to the low-frequency component and another corresponding to the low-frequency component of the stimulus. Fig.3 plots the variation of mean perceived polar elevation for low-frequency component, as a function of ITD variation-based elevation ΦΔITD and for three spectra-based elevation Φ_{spe}. The results for the low-frequency component are similar to those of 3 kHz low-pass filtered stimulus and can be summarized as follows:

(1) The front-back location of low-frequency virtual sources basically follows that of $\Phi_{\Delta ITD}$. The percentage η_{FB} of front-back confusion referred to $\Phi_{\Delta ITD}$ do not exceeds 10% in most cases. For a few exceptions, η_{FB} lies between 10% to 25%. Only for one case of ($\Phi_{\Delta ITD}$, Φ_{spe}, η_{FB}) = (120º, 0º, 37.5%), η_{FB} exceeds a limit of 25%.

(2) The up-down locations of perceived virtual source exhibit some confusion. The percentages η_{UD} of up-down confusion referred to $\Phi_{\Delta ITD}$ vary from 0% to a high value of 81.3%.

(3) After resolving the front-back and up-down confusion, the mean perceived elevation varies basically consistent with $\Phi_{\Delta ITD}$, in spite of Φ_{spe}.

In contrast, the results for the high-frequency component of pink noise can be summarized as follows:

(1) The front-back location of high-frequency virtual sources basically follows that of Φ_{spe}. The percentage η_{FB} of front-back confusion referred to Φ_{spe} is less than 10% in most case. For a few exceptions, η_{FB} lies between 10% to 15.6%. Only for one case of (Φ_{spe}, $\Phi_{\Delta ITD}$, η_{FB}) = (60º, 90º, 28.1%), η_{FB} exceeds a limit of 25%.

(2) The up-down locations of perceived virtual source exhibit low confusion. The percentages η_{UD} of up-down confusion referred to Φ_{spe} vary from 0% to 18.8%.

(3) After resolving the front-back and up-down confusion, the mean perceived elevation for the high-frequency component varies basically consistent with Φ_{spe}, in spite of $\Phi_{\Delta ITD}$.

The above observation can be further proved by applying an ANOVA to the experimental results.

5. DISCUSSION

The results in present experiment prove that the dynamic variation of ITD caused by head rotation dominates the front-back localization at low frequency, which is consistent with some previous results [6-7]. Two exceptions for η_{FB} more than 25% occur at high ITD variation-based elevation with $\Phi_{\Delta ITD}$ = 60º or 120º. The reason may be that at these high $\Phi_{\Delta ITD}$, the ITD variation caused by head rotation is small and then difficult to be perceived [See Fig.1(a)]. Another exceptions with η_{FB} = 30% occur at low ITD variation-based elevation with $\Phi_{\Delta ITD}$ = 210º. The reason is needed to be further explored.

Moreover, large variation in the percentages of up-down confusion at low frequency may be due to the fact that individualized HRTFs without torso were used in the experiment. The low-frequency spectral cue provided by torso, which may be important cues for up-down discrimination in addition to dynamic cue, was eliminated in the experiment.

On the other hand, the spectral cue dominates the front-back localization at high-frequency and only a few front-back confusions exist. The spectral cue also dominates the up-down discrimination at high frequency.

The results in present experiment also prove that dynamic and spectral cues dominate the vertical localization (perception of vertical displacement from the horizontal plane) at low frequency below 3 kHz and high frequency above 3 kHz, respectively. Our previous work indicated that the information providing by both cues is somewhat redundant [9]. When one cue is eliminated (but not conflicting), another cue alone still enables vertical localization to some extent. The experiment in present work further proves that conflicting dynamic and spectral cues result in two splitting virtual sources at different directions. This is different from the hypothesis in some previous work that spectra cue dominates vertical localization [1-4].

6. CONCLUSION

Both dynamic and spectral cues contribute to auditory front-back and vertical localization. The dynamic cue dominates front-back and vertical localization at low frequency below 3 kHz, and the spectral cue dominates front-back and vertical localization at high frequency above 3 kHz. One cue alone still enables vertical localization to some extent. But conflicting spectral and dynamic cues results in two splitting perceived virtual sources. The vertical positions of low-frequency and high-frequency virtual source are dominated by the dynamic and spectral

cues, respectively. It should be pointed out that the conclusion in present work is suitable for the pink noise. For other wideband stimuli, further validations are needed.

7. ACKNOWLEDGEMENTS

This study was supported by the National Natural Science Foundation of China (11674105) and the State Key Lab of Subtropical Building Science, South China University of Technology.

8. REFERENCES

[1] Blauert J: *Spatial Hearin*, Revised edition, MIT Press, Cambridge, MA, England, 1997.

[2] Xie BS: *Head-related transfer function and virtual auditory display*, J Ross Publishing, 2013.

[3] Hebrank J, and Wright D: "Spectral cues used in the localization of sound sources on the median plane," *J Acoust Soc Am*, 56 (6):1829–1834, 1974.

[4] Butler RA, and Belendiuk K.: "Spectral cues utilized in the localization of sound in the median sagittal plane". *J Acoust Soc Am*, 61(5):1264–1269, 1977.

[5] Wallach H.: "The role of head movement and vestibular and visual cue in sound localization," *J.Exp.Psychol*. 27 (4), 339-368, 1940.

[6] Wightman FL, and Kistler DJ.: "Resolution of front-back ambiguity in spatial hearing by listener and source movement," *J Acoust Soc Am* ,105(5):2841–2853, 1999.

[7] Macpherson EA.: "Head motion, spectral cues and Wallach's 'principle of least displacement' in sound localization.," *Principles and applications of spatial hearing*, 103–120, 2011.

[8] Perrett S, Noble W.: "The effect of head rotations on vertical plane sound localization," *J Acoust Soc Am* , 102(4):2325–2332, 1997.

[9] Jiang J L, Xie B S, Mai H M, et al.: "The role of dynamic cue in auditory vertical localization," *Appl. Acoust*., 146: 398-408, 2019.

[10] Y.Q. Rui, G.Z. Yu, B.S. Xie, et al.: "Calculation of individualized near-field head-related transfer function database using boundary element method," *in Audio Engineering Society Convention 134*, 2013.

[11] Kulkarni A., Isabelle S.K., and Colburn H.S. "Sensitivity of human subjects to head-related transfer-function phase spectra," *J. Acoust. Soc. Am.* 105(5), 2821-2840, 1999.

[12] Zhang C Y, Xie B S.: "Platform for dynamic virtual auditory environment real-time rendering system," *Chin. Sci. Bul.*, 58(3): 316-327, 2013.

 doi:10.25836/sasp.2019.06

TOWARDS A PERCEPTUALLY OPTIMAL BIAS FACTOR FOR DIRECTIONAL BIAS EQUALISATION OF BINAURAL AMBISONIC RENDERING

Thomas McKenzie, Damian T. Murphy, Gavin Kearney
AudioLab, Communication Technologies Research Group,
Department of Electronic Engineering, University of York,
York, YO10 5DD, UK
thomas.mckenzie@york.ac.uk

ABSTRACT

In virtual reality applications, where Ambisonic audio is presented to the user binaurally (over headphones) in conjunction with a head-mounted display, retained immersion relies on congruence between the auditory and visual experiences. Therefore frontal accuracy of binaural Ambisonic audio is an area of interest. A previously introduced method for improving the frequency reproduction of Ambisonic binaural rendering of a specific direction, called directional bias equalisation, is here applied to higher order Ambisonics and evaluated both numerically and perceptually. This paper attempts to obtain the optimal amount of directional bias, to find the best compromise between improved frontal reproduction and reduced lateral reproduction accuracy.

1. INTRODUCTION

Ambisonics is a method of generating, capturing and rendering two or three-dimensional sound fields, first developed by Michael Gerzon in the 1970s [1]. It has enjoyed a recent resurgence in popularity due to virtual reality applications, where Ambisonic audio can be presented to the user binaurally [2, 3] in conjunction with a head-mounted display, due to the relative ease of Ambisonic soundfield rotation. In virtual reality scenarios however, it is imperative to maximise the coherence between audio in the frontal direction and visuals in order to maintain immersion.

Ambisonic reproduction can theoretically be perfect in the centre of the loudspeaker array for frequencies up to the 'spatial aliasing frequency', f_{alias}. At frequencies above f_{alias} however, the limited spatial accuracy of reproducing a physical sound field with a finite number of transducers produces artefacts such as localisation blur, reduced lateralisation and comb filtering. Increasing the order of Ambisonics delivers a more accurate soundfield reconstruction, but requires a greater number of channels and convolutions in binaural rendering. One approach for improving spectral reproduction of binaural Ambisonic rendering is diffuse-field equalisation [4]. In a previous study, the authors applied this technique to virtual loudspeaker binaural Ambisonic decoders, which was shown to improve both the spectral response over the sphere and predicted median plane elevation localisation. However, it was evident there still exists a definite and perceivable difference in timbre between diffuse-field equalised binaural Ambisonic rendering and HRTF convolution, even at 5th order Ambisonics.

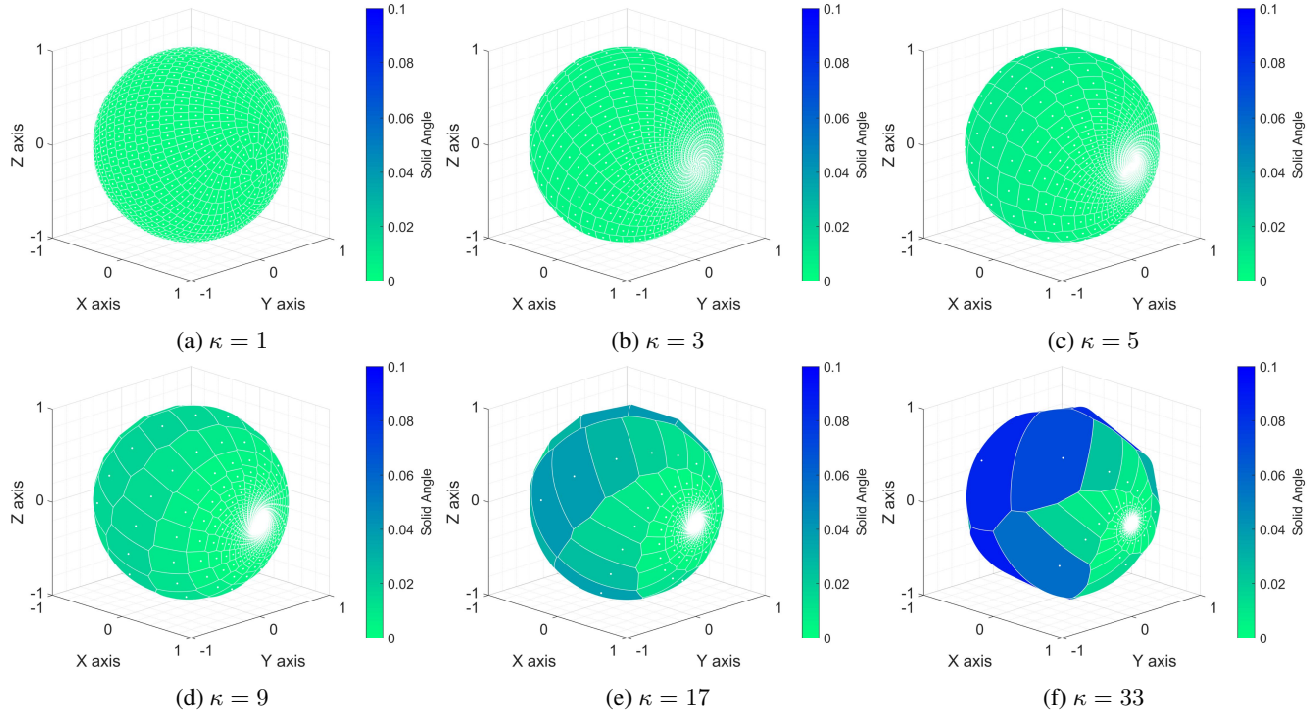
Altering the diffuse-field equalisation method by concentrating the equalisation in one specific direction has been shown to be possible by creating a hybrid of free-field and diffuse-field equalisation. Reproduction in the desired direction can then be made more accurate at the expense of other directions, such that an infinite directional bias can theoretically produce binaural Ambisonic audio equivalent to HRTF convolution. This is referred to as directional bias equalisation (DBE) [5]. DBE is a pre-processing stage that can be implemented offline.

In the initial paper detailing the DBE method, evaluation focused on application to 1st order Ambisonics with only a numerical analysis presented. This paper follows on from the first paper, extending the evaluation of the technique to higher-order Ambisonics, with both a numerical and perceptual evaluation. Here 1st, 3rd and 5th order Ambisonics are investigated, with loudspeaker configurations comprising 6, 26 and 50 loudspeakers respectively, arranged in Lebedev grids [6] (for exact vertices, see [7]). The 6 values of the bias factor κ used in this paper are: $\kappa = 1, 3, 5, 9, 17$ and 33. Though DBE can increase the fidelity of any desired direction, this paper will focus on frontal bias at a direction of $(\theta, \phi) = (0°, 0°)$, where θ and ϕ denote azimuth and elevation, respectively.

In this study Ambisonics was rendered binaurally with mode-matching pseudo-inverse decoding using the Politis Ambisonic library [8]. Three-dimensional full normalisation (N3D) and Ambisonic channel number (ACN) ordering was used throughout. Dual-band decoding was implemented with Max r_E weighting [9, 10] at frequen-

© Thomas McKenzie, Damian T. Murphy, Gavin Kearney. Licensed under a Creative Commons Attribution 4.0 International License (CC BY 4.0). **Attribution:** Thomas McKenzie, Damian T. Murphy, Gavin Kearney. "Towards a perceptually optimal bias factor for directional bias equalisation of binaural Ambisonic rendering", 1st EAA Spatial Audio Signal Processing Symposium, Paris, France, 2019.

doi:10.25836/sasp.2019.08



(a) $\kappa = 1$ (b) $\kappa = 3$ (c) $\kappa = 5$

(d) $\kappa = 9$ (e) $\kappa = 17$ (f) $\kappa = 33$

Figure 1: Voronoi spheres demonstrating the directionally biased quadratures used in the first stage of DBE for the 6 values of κ used in this study.

cies above the spatial aliasing frequency, which in this study was approximated according to [11, 12] as 670 Hz, 1870 Hz and 3070 Hz for 1st, 3rd and 5th order Ambisonics, respectively, calculated with a speed of sound of 343 m/s and radius of the listening area as 9 cm (the approximate radius of the Neumann KU 100 dummy head). All head-related impulse responses (HRIRs) used in this study were from the Bernschütz Neumann KU 100 database [13]. All computation was carried out offline in MATLAB version 9.3.0 - R2017b. All audio used was of 24-bit depth and 48 kHz sample rate.

2. DIRECTIONAL BIAS EQUALISATION

DBE is a two stage equalisation process. For a detailed explanation of the DBE method, the reader is directed to the original paper [5]. Here the method will be briefly summarised. Firstly, a directionally biased quadrature (DBQ) RMS response is calculated from the root-mean-square (RMS) of the magnitude responses of a large number of binaural Ambisonic renders, taken at locations over the sphere. The quadrature is directionally biased by skewing the z axis values according to κ such that:

$$z_\beta = \kappa(z_\alpha + 1) - 1 \tag{1}$$

The cartesian coordinates are then converted back to spherical coordinates and rotated to the direction of bias. In this study, Fibonacci quadrature with 1000 points is used in DBQ RMS calculations, due to its relative even distribution of points. The Voronoi sphere plots of the 6 values of κ used in this paper are shown in Figure 1. With no directional bias (see Figure 1a), points are evenly distributed over the sphere, which when equalised produces an even RMS response of the binaural Ambisonic loudspeaker configuration, theoretically equivalent to diffuse-field equalisation [4].

The second stage of DBE is a directional bias HRTF equalisation. In this study the direction of bias is chosen as directly in front: $(\theta, \phi) = (0°, 0°)$. As κ increases, the gain of the frontal bias HRTF g_β increases such that:

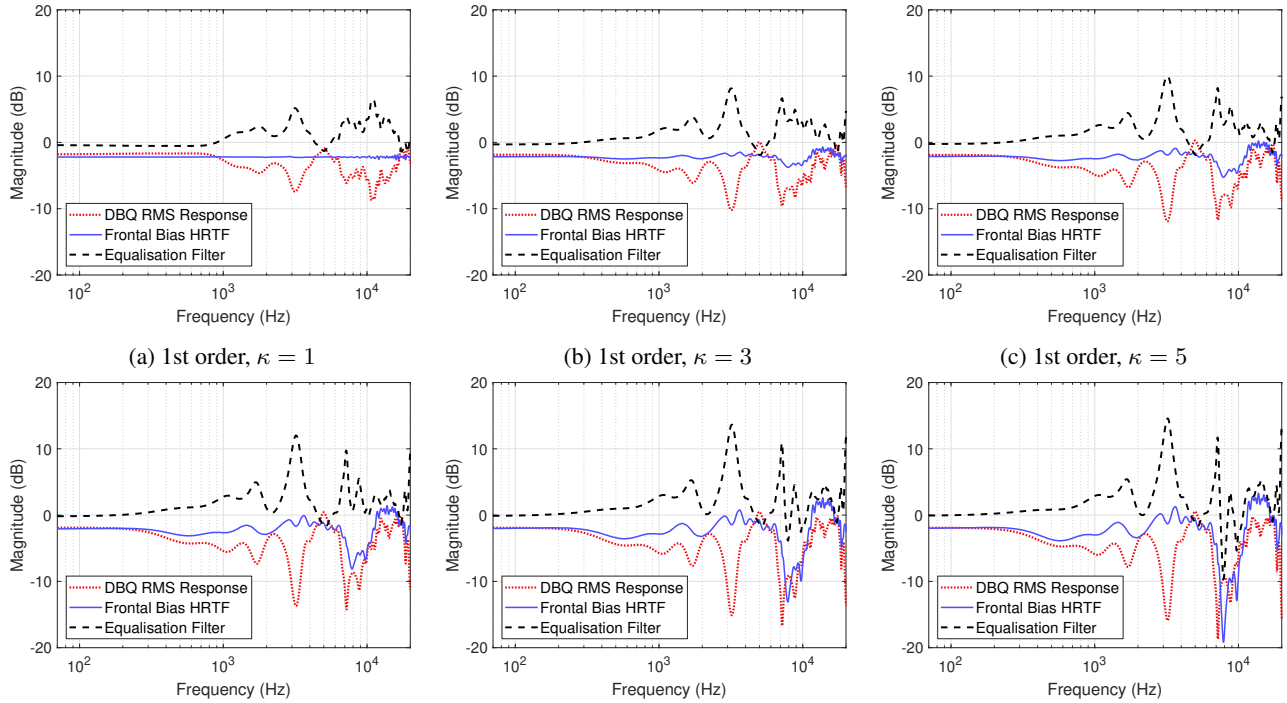
$$g_\beta = 1 - e^{-\frac{\kappa - 1}{10}} \tag{2}$$

The frontal bias HRTF filter is then convolved with the DBQ RMS equalisation resulting in the final DBE filters. Figure 2 presents the DBQ RMS responses, frontal bias HRTFs, and resulting overall equalisation filters for 1st order, with varying κ. To conserve space, 3rd and 5th order plots have been omitted. As κ increases, the DBQ response closer resembles a frontal Ambisonic render, and the frontal bias HRTF closer resembles a frontal HRTF.

3. NUMERICAL EVALUATION

To assess the effect of varying directional bias on the spectral accuracy of binaural Ambisonic signals, a perceptually motivated fast Fourier transfer (FFT) based spectral difference (PSD) model was used that takes into account various features of human auditory perception [14]. The PSD model weights input signals using ISO 226 equal loudness contours [15] to account for the frequency-varying sensitivity of human hearing, uses a sone scale to account for the loudness-varying sensitivity of human hearing, and equivalent rectangular bandwidth weightings to address how the

doi:10.25836/sasp.2019.08



(a) 1st order, $\kappa = 1$ (b) 1st order, $\kappa = 3$ (c) 1st order, $\kappa = 5$

(d) 1st order, $\kappa = 9$ (e) 1st order, $\kappa = 17$ (f) 1st order, $\kappa = 33$

Figure 2: Directionally biased quadrature responses, frontal bias HRTF equalisation and resulting overall equalisation filters for 1st order Ambisonics with varying κ.

linearly spaced samples of an FFT do not accurately represent the approximately logarithmic frequency sensitivity of the inner ear.

PSD between binaural Ambisonic renders and HRIRs was calculated for all 3 tested orders of Ambisonics over 16,020 locations on the sphere, distributed using a 2° Gaussian grid. Figure 3 shows the solid angle weighted PSD value for each value of κ for all tested orders of Ambisonics, with whiskers denoting the maximum and minimum PSD values. It is clear that increasing κ reduces the minimum PSD value whilst increasing the maximum PSD value. At $\kappa = 33$, the original HRTF is almost perfectly reconstructed, even with 1st order Ambisonics. However an interesting trend appears to suggest that with increased Ambisonic order, the value of κ needs to be greater to achieve a similar minimum PSD. To observe how PSD changes over the sphere with varying κ, 1st order PSD plots are presented in Figure 4 (to reduce the overall amount of figures, 3rd and 5th order plots were omitted). This shows how increasing the value of κ produces an improvement in spectral accuracy for the frontal direction with a reduction in lateral accuracy, which is in line with expectations. 3rd and 5th order renders also follow this pattern.

4. PERCEPTUAL EVALUATION

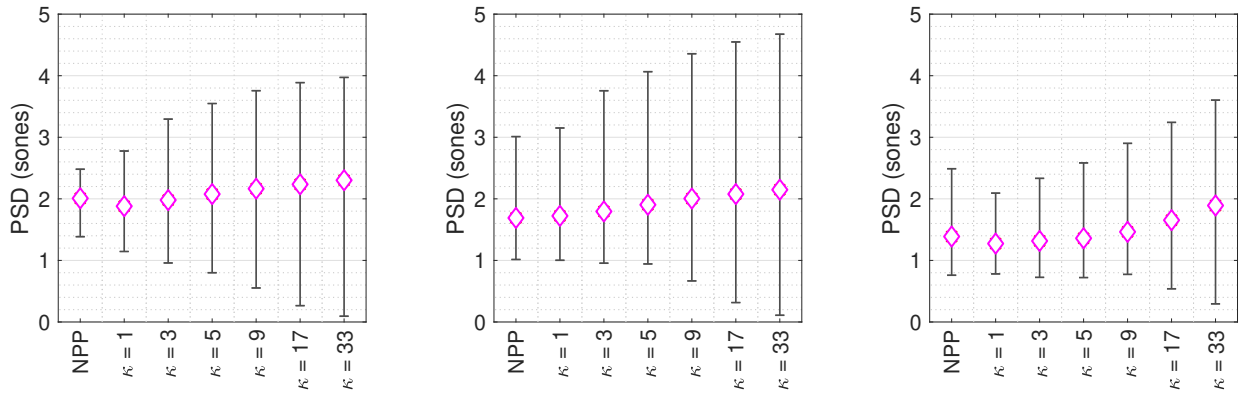
To assess the perceptual effect of varying κ, listening tests were conducted using both simple and complex acoustic scenes. Tests followed the multiple stimulus with hidden reference and anchors (MUSHRA) paradigm, ITU-R BS.1534-3 [16], and were conducted in a quiet listening room using a single set of Sennheiser HD 650 circum-aural headphones and an Apple Macbook Pro with a Fireface 400 audio interface, which has software controlled input and output levels. The headphones were equalised using a Neumann KU 100 from the RMS average of 11 impulse response measurements collected using Farina's swept sine technique [17]. Inverse filtering was achieved using Kirkeby and Nelson's least-mean-square regularization method [18], with one octave smoothing implemented using the complex smoothing approach of [19] and an inversion range of 5 Hz–4 kHz. In-band and out-band regularization of 25 dB and −2 dB respectively was used to avoid narrow peaks in the inverse filter, which are more noticeable than notches [20]. 20 experienced listeners participated, aged between 22 and 41, with no reported knowledge of any hearing impairments.

4.1 Test Methodology

The simple scenes test comprised of a single pink noise source. Two sound source locations were used: directly in front of the listener at $(\theta, \phi) = (0°, 0°)$, and directly to the left of the listener at $(\theta, \phi) = (90°, 0°)$. The reference was a direct HRIR convolution, and low and mid anchors were the reference low-passed at 3.5 kHz and 7 kHz, respectively.

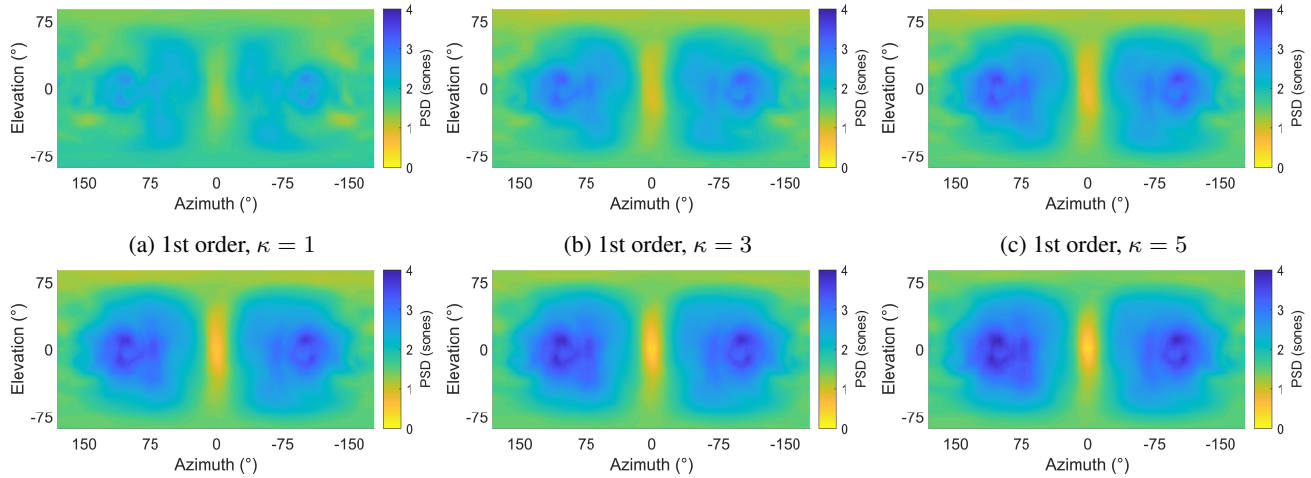
The complex scene was simulated by mixing a pink noise burst with a diffuse soundscape. The noise burst consisted of 0.5s burst followed by 0.5s of silence panned directly in front of the listener. The diffuse soundscape was synthesised from 24 excerpts of a monophonic sound scene

doi:10.25836/sasp.2019.08



(a) 1st order (b) 3rd order (c) 5th order

Figure 3: Solid angle weighted perceptual spectral difference between 16,020 binaural Ambisonic renders and HRIRs with varying κ, for 1st, 3rd and 5th order Ambisonics. Whiskers denote maximum and minimum PSD values and NPP denotes no pre-processing.



(a) 1st order, $\kappa = 1$ (b) 1st order, $\kappa = 3$ (c) 1st order, $\kappa = 5$

(d) 1st order, $\kappa = 9$ (e) 1st order, $\kappa = 17$ (f) 1st order, $\kappa = 33$

Figure 4: Perceptual spectral difference between 1st-order binaural Ambisonic renders and HRIRs over the sphere with varying κ.

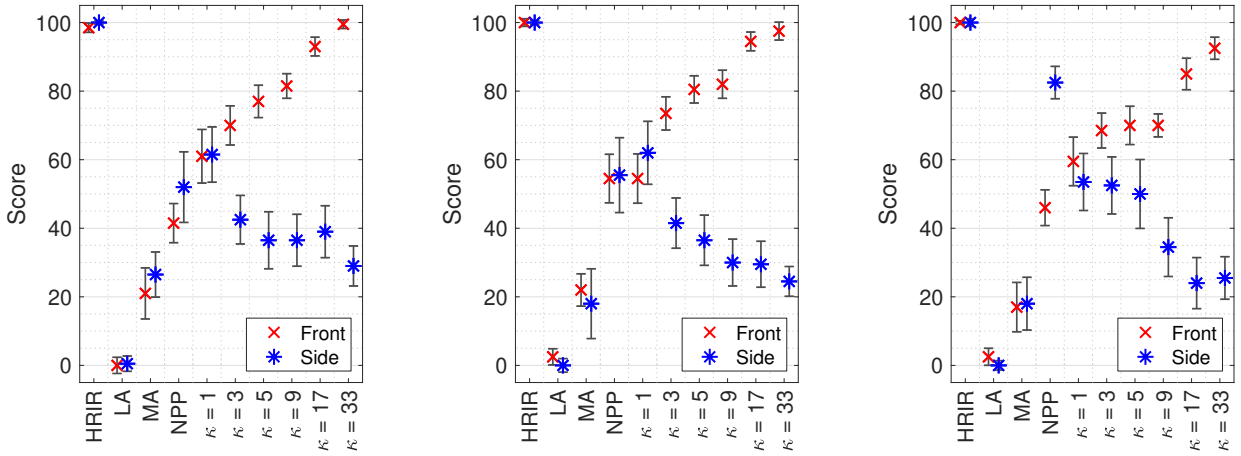
recording of a train station [21], each 5s long. The sound scene excerpts were panned to the vertices of a spherical 24 pt. T-design quadrature [22], to ensure minimal overlap between virtual loudspeaker positions in the binaural decoders and the sound sources in the complex scene. The frontal noise was set as 3 dB RMS louder than the diffuse soundscape to approximate a centre of attention. The reference comprised of a sum of direct HRIR convolutions and the anchor a 0th order Ambisonic render. All test scenarios were repeated once.

4.2 Results

Results data was tested for normality using the one-sample Kolmogorov-Smirnov test, which showed all data as non-normal. Results were therefore analysed using non-parametric statistics. Figure 5 presents the simple scene median scores for orders 1st, 3rd and 5th order Ambisonics with non-parametric 95% confidence intervals [23]. A Friedman's ANOVA, conducted on simple scene data from all orders and sound source locations, confirmed that changing the value of κ had a statistically significant effect on the perceived similarity to the HRTF reference ($\chi^2(6) = 27.22, p < 0.01$). The results support the theory that increasing κ improves the perceived similarity to the HRTF reference for the frontal stimuli for all 3 tested orders of Ambisonics, and reduces the similarity for the lateral stimuli. This shows that DBE performs as expected with simple scenes.

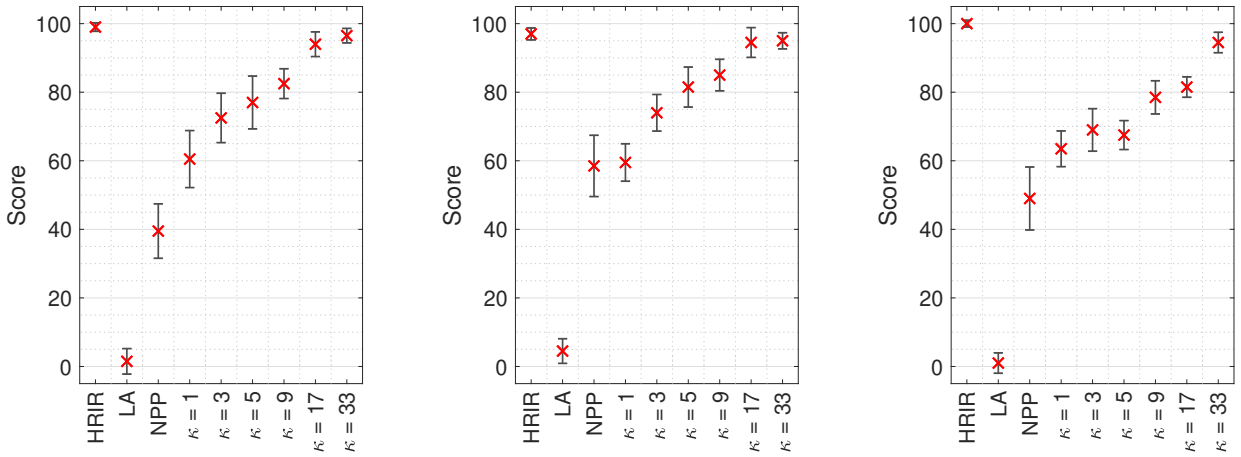
Figure 6 presents the complex scene median scores for 1st, 3rd and 5th order Ambisonics with non-parametric 95% confidence intervals [23]. A Friedman's ANOVA, conducted on complex scene data from all orders, again confirmed that changing the value of κ had a statistically significant effect on the perceived similarity to the HRTF reference ($\chi^2(6) = 383.47, p < 0.01$). Interestingly, the results suggest that the diffuse sound was essentially ignored, as results for the complex scene are similar to the frontal stimuli in the simple scene, with increasing κ producing a higher perceived similarity to the HRTF references for all 3 tested orders of Ambisonics.

doi:10.25836/sasp.2019.08



(a) 1st order (b) 3rd order (c) 5th order

Figure 5: Median scores of the simple scene tests with non-parametric 95% confidence intervals. Scores indicate perceived similarity to the HRIR reference. LA, MA and NPP denote low anchor, medium anchor and no pre-processing, respectively.



(a) 1st order (b) 3rd order (c) 5th order

Figure 6: Median scores of the complex scene tests with non-parametric 95% confidence intervals. Scores indicate perceived similarity to the HRIR reference. LA and NPP denote low anchor and no pre-processing, respectively.

5. DISCUSSION AND CONCLUSIONS

This paper has presented a numerical and perceptual evaluation of directional bias equalisation for 1st, 3rd and 5th order binaural Ambisonic rendering. It has shown how, as directional bias increases, the spectral performance in the direction of bias improves, but to the detriment of other directions. However when the scene is complex, with a frontal main source and additional diffuse sources, increased frontal bias still improves the performance. This suggests that, if the main sound source is at the front, one can afford to increase κ without greatly reducing the perceived quality in lateral directions. However, if the stimuli is more diffuse, the value of κ should be reduced.

Future work will look at combining with other pre-processing techniques for improving binaural Ambisonic reproduction, including ILD optimisation [24] and time alignment [25–27]. As both have been shown to improve spectral reproduction, this could mean that a lower value of κ is required to get perceptually equivalence to HRIR convolution in the direciton of bias. Additionally, if the dominant direction of the signal can be estimated (for example using a method such as directional audio coding [28]), the equalisation could be performed specifically for the direction of arrival, such as other signal dependent Ambisonic decoding methods [29, 30].

6. REFERENCES

[1] M. A. Gerzon, "Periphony: with-height sound reproduction," *Journal of the Audio Engineering Society*, vol. 21, no. 1, 1973.

[2] J.-M. Jot, S. Wardle, and V. Larcher, "Approaches to binaural synthesis," in *105th Convention of the Audio Engineering Society*, vol. 105, 1998.

[3] M. Noisternig, A. Sontacchi, T. Musil, and R. Höldrich, "A 3D Ambisonic based binaural sound reproduction system," in *AES 24th International Conference on Multichannel Audio*, 2003.

[4] T. McKenzie, D. Murphy, and G. Kearney, "Diffuse-field equalisation of binaural Ambisonic rendering," *Applied Sciences*, vol. 8, no. 10, 2018.

[5] T. McKenzie, D. Murphy, and G. Kearney, "Directional bias equalisation of first-order binaural Ambisonic rendering," in *AES Conference on Audio for Virtual and Augmented Reality*, 2018.

[6] V. I. Lebedev, "Quadratures on a sphere," *USSR Computational Mathematics and Mathematical Physics*, vol. 16, no. 2, pp. 10–24, 1976.

[7] J. Burkardt, "SPHERE_LEBEDEV_RULE - Quadrature Rules for the Unit Sphere." 2013.

[8] A. Politis, *Microphone array processing for parametric spatial audio techniques*. PhD thesis, Aalto University, 2016.

[9] M. A. Gerzon and G. J. Barton, "Ambisonic decoders for HDTV," in *92nd Convention of the Audio Engineering Society*, 1992. Preprint 3345.

[10] J. Daniel, J.-B. Rault, and J.-D. Polack, "Ambisonics encoding of other audio formats for multiple listening conditions," in *105th Convention of the Audio Engineering Society*, 1998. Preprint 4795.

[11] S. Moreau, J. Daniel, and S. Bertet, "3D sound field recording with higher order Ambisonics - objective measurements and validation of a 4th order spherical microphone," in *120th Convention of the Audio Engineering Society*, 2006.

[12] S. Bertet, J. Daniel, E. Parizet, and O. Warusfel, "Investigation on localisation accuracy for first and higher order Ambisonics reproduced sound sources," *Acta Acustica united with Acustica*, vol. 99, no. 4, pp. 642–657, 2013.

[13] B. Bernschütz, "A spherical far field HRIR/HRTF compilation of the Neumann KU 100," in *Fortschritte der Akustik – AIA-DAGA 2013*, pp. 592–595, 2013.

[14] C. Armstrong, T. McKenzie, D. Murphy, and G. Kearney, "A perceptual spectral difference model for binaural signals," in *145th Convention of the Audio Engineering Society*, 2018. Convention E–Brief 457.

[15] International Organization for Standardization, "ISO 226:2003, Normal equal-loudness-level contours," 2003.

[16] ITU-R-BS.1534-3, "Method for the subjective assessment of intermediate quality level of audio systems BS Series Broadcasting service (sound)," vol. 3, 2015.

[17] A. Farina, "Simultaneous measurement of impulse response and distortion with a swept-sine technique," in *108th Convention of the Audio Engineering Society*, 2000. Preprint 5093.

[18] O. Kirkeby and P. A. Nelson, "Digital filter design for inversion problems in sound reproduction," *Journal of the Audio Engineering Society*, vol. 47, no. 7/8, pp. 583–595, 1999.

[19] P. D. Hatziantoniou and J. N. Mourjopoulos, "Generalized fractional-octave smoothing of audio and acoustic responses," *Journal of the Audio Engineering Society*, vol. 48, no. 4, pp. 259–280, 2000.

[20] R. Bücklein, "The audibility of frequency response irregularities," *Journal of the Audio Engineering Society*, vol. 29, no. 3, pp. 126 – 131, 1981.

[21] M. Green and D. Murphy, "EigenScape: a database of spatial acoustic scene recordings," *Applied Sciences*, vol. 7, no. 12, 2017.

[22] R. H. Hardin and N. J. A. Sloane, "McLaren's improved Snub cube and other new spherical designs in three dimensions," *Discrete & Computational Geometry*, vol. 15, no. 4, pp. 429–441, 1996.

[23] R. Mcgill, J. W. Tukey, and W. A. Larsen, "Variations of box plots," *The American Statistician*, vol. 32, no. 1, pp. 12–16, 1978.

[24] T. McKenzie, D. Murphy, and G. Kearney, "Interaural level difference optimisation of binaural Ambisonic rendering," *Applied Sciences*, vol. 9, no. 6, 2019.

[25] M. J. Evans, J. A. S. Angus, and A. I. Tew, "Analyzing head-related transfer function measurements using surface spherical harmonics," *Journal of the Acoustical Society of America*, vol. 104, no. 4, pp. 2400–2411, 1998.

[26] J. G. Richter, M. Pollow, F. Wefers, and J. Fels, "Spherical harmonics based hrtf datasets: Implementation and evaluation for real-time auralization," *Acta Acustica united with Acustica*, vol. 100, no. 4, pp. 667–675, 2014.

[27] M. Zaunschirm, C. Schörkhuber, and R. Höldrich, "Binaural rendering of Ambisonic signals by HRIR time alignment and a diffuseness constraint," *Journal of the Acoustical Society of America*, vol. 143, no. 6, pp. 3616–3627, 2018.

[28] V. Pulkki, "Spatial sound reproduction with directional audio coding," *Journal of the Audio Engineering Society*, vol. 55, no. 6, pp. 503–516, 2007.

[29] A. Politis, L. McCormack, and V. Pulkki, "Enhancement of ambisonic binaural reproduction using directional audio coding with optimal adaptive mixing," *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, vol. 2017, pp. 379–383, 2017.

[30] C. Schörkhuber and R. Höldrich, "Linearly and quadratically constrained least-squares decoder for signal-dependent binaural rendering of Ambisonic signals," in *AES Conference on Immersive and Interactive Audio*, (York), 2019. Paper 22.

 doi:10.25836/sasp.2019.08

THE SPHEAR PROJECT UPDATE: REFINING THE OCTASPHEAR, A 2ND ORDER AMBISONICS MICROPHONE

Fernando Lopez-Lezcano
CCRMA, Stanford University
nando@ccrma.stanford.edu

ABSTRACT

This paper presents an update of the *SpHEAR (Spherical Harmonics Ear) project, created with the goal of using low cost 3D printers to fabricate Ambisonics microphones. The project includes all mechanical 3d models and electrical designs, as well as all the procedures and software needed to calibrate the microphones. Everything is shared through GPL/CC licenses and is available in a public GIT repository. [1] We will focus on the status of the eight-capsule OctaSpHEAR 2nd order microphone, with details of the evolution of its mechanical design and calibration.

1. INTRODUCTION

The soundfield microphone was designed in the 1970s by Michael Gerzon and Peter Craven [1] to capture the spherical harmonics of a soundfield up to first-order. It uses four capsules in a tetrahedral configuration, which are matrixed and equalized to derive the Ambisonics B-format signals that represent the soundfield. In 2012 Eric Benjamin published the design and preliminary evaluation of an eight capsule microphone [2], which can capture second order Ambisonics components and shows better performance in first order than the traditional tetrahedral microphone. Its capsules are located in the vertices of a square antiprism, and it can encode 8 of the 9 components of an Ambisonics 2nd order soundfield (figure 1). The R component cannot be recovered as it aliases to W. This design is the basis of our OctaSpHEAR (aka: Octathingy) microphone.

The SpHEAR project started at the end of 2015 with the design and construction of conventional tetrahedral prototypes [3]. These initial designs were followed by eight capsule prototypes [4], with a calibration procedure derived from the work on the tetrahedral microphones. This paper focuses primarily on the eight-capsule design. It presents two different acoustical and mechanical designs of the capsule array, and compares their raw and calibrated performance. It also explore optimizations of the encoding process in the high frequency range.

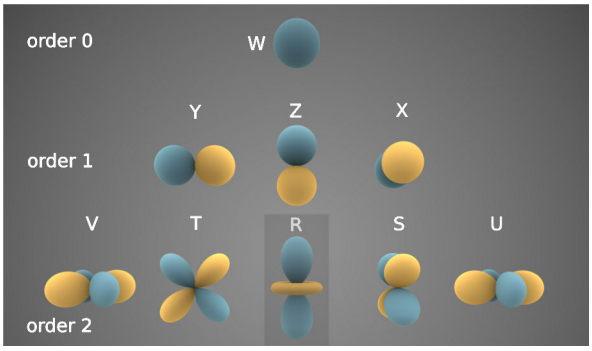


Figure 1. 2nd order spherical harmonics [2]

2. MECHANICAL DESIGN, VERSION 1 (V1)

The mechanical design of the OctaSpHEAR's first two prototypes was a direct derivation of the tetrahedral design. The capsule array is created out of individual capsule holders that assemble together like a 3D puzzle.



Figure 2. OctaSpHEAR v1 capsule array and individual capsule holder

The array was designed with a radius of 18mm, which is close to the minimum that can be obtained with 14mm diameter capsules.

The first two prototypes built have been extensively used for field recordings, and concert and event documentation at CCRMA, and their performance has been considered very adequate when compared to much more expensive microphones. Nevertheless, a plot of the raw frequency response of individual capsules in the array show

[1] https://cm-gitlab.stanford.edu/ambisonics/SpHEAR/
[2] https://commons.wikimedia.org/wiki/File:Spherical_Harmonics.png

© Fernando Lopez-Lezcano. Licensed under a Creative Commons Attribution 4.0 International License (CC BY 4.0). **Attribution:** Fernando Lopez-Lezcano. "The SpHEAR Project Update: refining the OctaSpHEAR,
a 2nd order Ambisonics microphone", 1st EAA Spatial Audio Signal Processing Symposium, Paris, France, 2019.

doi:10.25836/sasp.2019.34

problems that suggest a better design could improve the performance of the microphone.

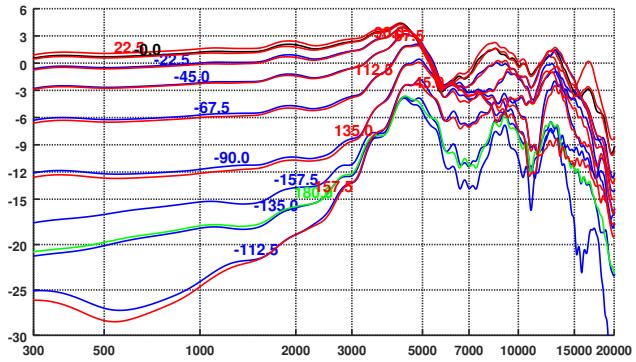


Figure 3. OctaSpHEAR v1 capsule #1 frequency response as a function of incident angle, as indicated on the corresponding trace (in degrees)

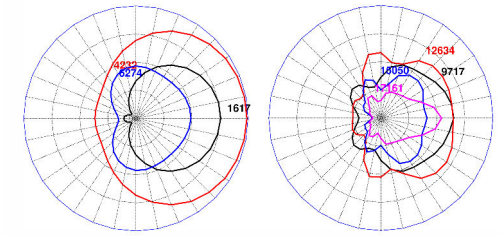


Figure 4. OctaSpHEAR v1 capsule polar patterns at different frequencies

The plots show a strong resonant peak at around 4.4KHz (and its harmonics) caused by the space enclosed by the eight capsules which creates a Helmholtz resonator with multiple necks. The resonances degrade the polar pattern of the capsule at frequencies above about 3KHz. The front to back ratio is reduced, and the capsules become more omnidirectional. This will introduce distortions in the shape of the recovered Ambisonics components.

3. MECHANICAL DESIGN, VERSION 2 (V2)

The resonances suggested a different approach (common to many existing commercial microphones) to the mechanical design of the array. The simple design was replaced by individual conical capsule holders that attach to a spherical core.

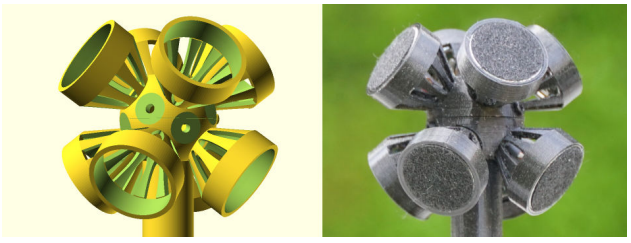


Figure 5. OctaSpHEAR v2 capsule array design

Mechanical design constraints forced us to use a bigger array radius than in version 1 (20.5mm instead of 18mm).

If we only attach one capsule holder to the version 2 design we can measure an almost ideal free field capsule response that still includes the effect of the capsule holder and the body of the microphone. This set of measurements helps us define a baseline performance for this capsule (Primo EMM200), and will help us understand how the rest of the microphone affects its performance.

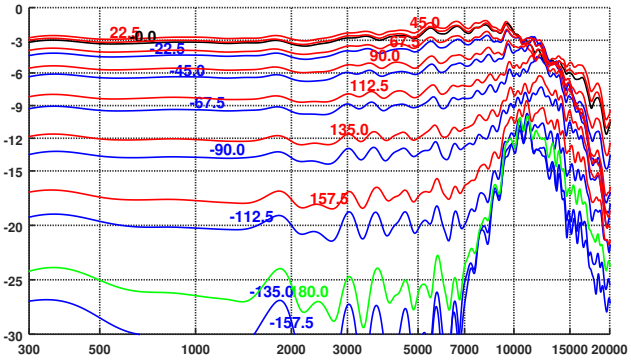


Figure 6. OctaSpHEAR v2, one capsule holder and capsule, frequency response as a function of incident angle

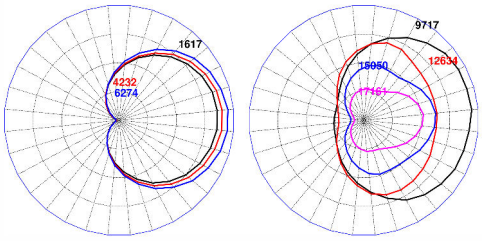


Figure 7. OctaSpHEAR v2, one capsule holder and capsule, polar pattern at different frequencies

Up to about 7Khz the capsule behaves almost like a perfect cardioid, above that we see a degradation of the polar pattern (figure 7) and it becomes more omnidirectional (an expected behavior in cardioid capsules).

Adding the other seven capsules changes the response as shown in figures 8 and 9.

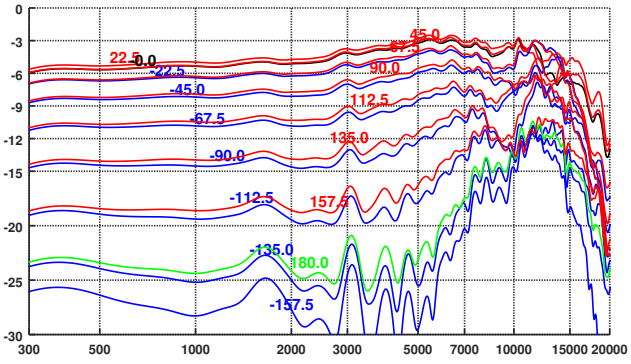


Figure 8. OctaSpHEAR v2 capsule #1 frequency response as a function of incident angle

The occlusion created by all the other capsules degrades the front to back ratio at low and mid frequencies, compared to the measurements of a single capsule. Even then,

doi:10.25836/sasp.2019.34

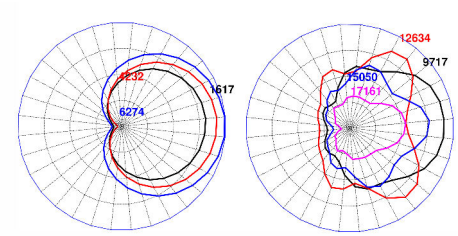


Figure 9. OctaSpHEAR v2 capsule #1 polar responses at different frequencies

the ratio is better than in the version 1 design, except at very low frequencies. The measurements confirm that the resonances at 4.4Khz are gone, as expected, and show that the polar patterns are more consistent over frequency.

In both designs, the polar patterns at very high frequencies in figure 4 and 9 show the shadowing effect of the other capsules and exhibit multiple lobes in their response.

4. MEASUREMENT AND CALIBRATION

As detailed in our previous paper [4], our microphones are measured using quasi-anechoic techniques, with an automated system based on a low cost modified robotic arm. The plots in this paper are derived from 32 equally spaced impulse response measurements in the horizontal plane and 150 measurements of an equally spaced 240 point spherical t-design [6] in the full sphere. Not all points in the t-design are reachable by the arm, which is currently limited due to its length to points lying between -24 and +54 degrees of elevation with respect to the horizontal plane. We obtain about 4.5mSecs of clean equalized impulse response data from each measurement.

The measured impulse responses are used to calibrate the microphone, that is, to create an encoder black box that converts the 8 capsule signals (A format) to 8 Ambisonics components (B format). In an ideal world the B format signals frequency response would be flat, they would be in phase over the full frequency range, and their polar patterns would match the theoretical ones and would not change over frequency.

A simple static 8×8 matrix cannot not satisfy these criteria as the spacing between capsules will create phase related boosts and cancellations in the B format signals above a transition frequency determined by the radius of the array. For our microphones this effect starts to show up at roughly 2KHz, and it can be mitigated by the design of suitable B format correction filters.

As shown in figure 6 the capsule itself does not behave like a cardioid at all frequencies. The polar plots in figure 7 shows it becomes a subcardioid and then a hypercardioid as frequency increases. The other capsules in the array and the structure that supports them also distorts its polar pattern, and the capsule show multiple lobes at very high frequencies (figures 4 and 9), as well as being more omnidirectional at low and mid frequencies in version 2.

These changes of directivity versus frequency will create frequency dependent distortions of the shape of the lobes of the recovered Ambisonics signals.

Finally, cardioid capsules have a peak in their on-axis response at high frequencies (for our capsules there is a 5dB boost at 12KHz, approximately, see figure 10). This peak gradually disappears at increasing angles of incidence from the axis of the capsule, and in our capsules the effect is almost gone at 35 degrees off-axis.

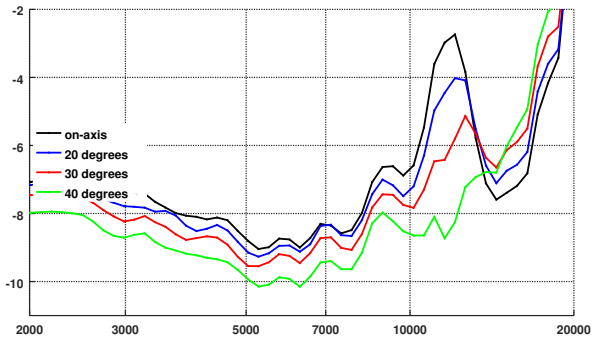


Figure 10. OctaSpHEAR v2, high frequency on-axis capsule resonance

Changes over frequency of the capsule polar patterns, and changes of the polar pattern that depend on the angle of incidence will create distortions in the recovered Ambisonics signals that we cannot really correct. We cannot fix capsule polar patterns, and all our processing and filtering is angle-invariant.

4.1 Encoding from A format to B format

In our current very simple encoder design strategy [4] we start by equalizing all capsules and then using singular value decomposition [7] to create an 8×8 static A to B conversion matrix in a range of frequencies where the capsules can be considered to be co-located. Using the capsule equalization filters and the A to B matrix, we create a first approximation of the B format signals, which deviates from theory above the transition frequency. These signals are then used to create B format equalization filters (figure 11) that mitigate those deviations.

For the horizontal components (WXYUV), the performance of the resulting encoder is different if we calculate it based on just horizontal plane measurements or all measurements. Horizontal only measurements yield better results (flatter frequency response, less variations at high frequencies). We choose to optimize the performance of the microphone for signals coming from the horizontal plane or from low positive or negative elevations, because this is the more likely source of sounds in “normal” recording conditions. So, the A to B matrix and B format correction filters for WXYUV are designed using horizontal plane measurements only. The encoder for the rest of the components (ZST) is designed using the full 3D set of measurements (we do not have a choice here as these are the measurements that have information about height). The final A to B matrix and B format correction filters are a merge of both.

Additionally, the second order components of the microphone are difference microphones, so their output drops at 6dB/octave throughout the full frequency range. The B format filters for those components include the needed

boost to equalize them, and we add regularization high pass filters that limit the capsule self-noise amplification.

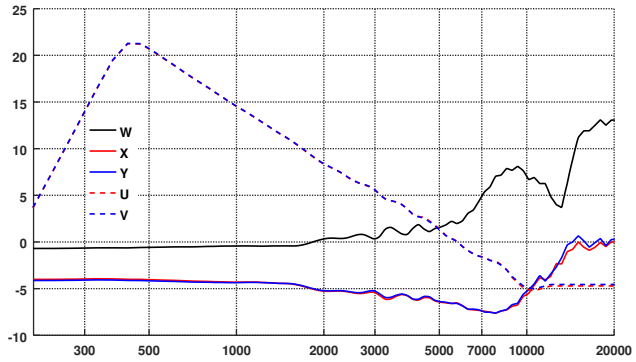


Figure 11. OctaSpHEAR v2 B format equalization filters

Even with these filters, the second order components are noticeably noisier than the first order components, so a set of defeatable expanders is included in the encoder to minimizes the noise for low level signals or silence.

4.2 Version 1 and 2, Ambisonics performance

Figures 12 through 19 show plots of the performance of the calibrated microphones. They include the frequency response of one 1st order (Y) and one 2nd order (V) Ambisonics signals in the horizontal plane, with the azimuth of the excitation signal as a parameter, and polar plots at different frequencies.

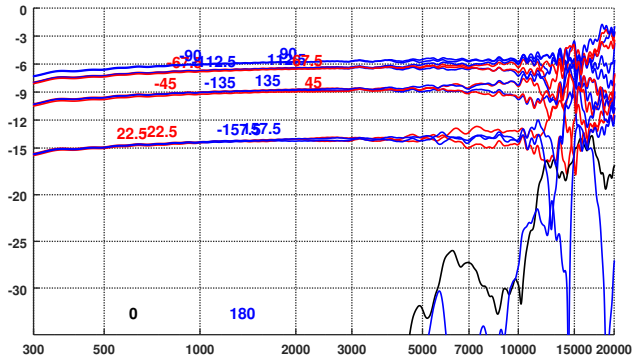


Figure 12. OctaSpHEAR v1 B format Y frequency response with the azimuth angle as a parameter

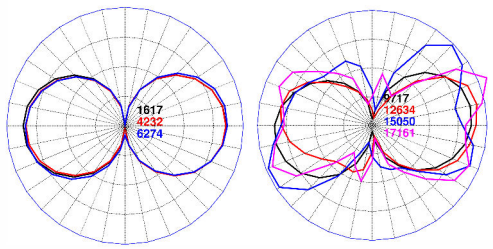


Figure 13. OctaSpHEAR v1 B format Y polar pattern

Both designs show very solid performance up to about 10-11Khz, with frequency response deviations from the ideal flat performance of only a few dB. Above that there is more spread of the response due to the changes in the polar patterns of the capsules.

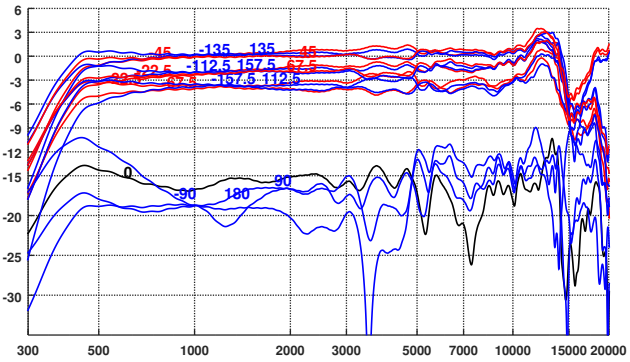


Figure 14. OctaSpHEAR v1 B format V frequency response with the azimuth angle as a parameter

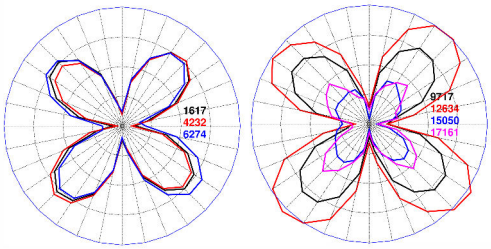


Figure 15. OctaSpHEAR v1 B format V polar pattern

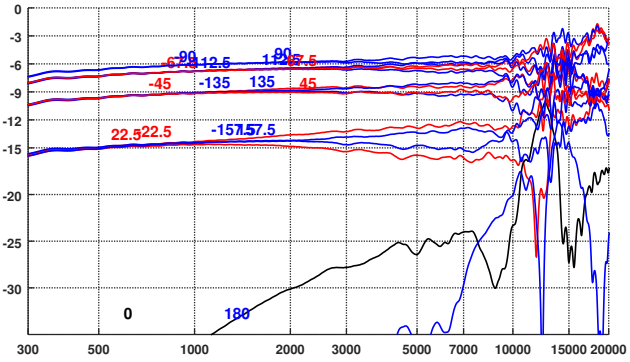


Figure 16. OctaSpHEAR v2 B format Y frequency response with the azimuth angle as a parameter

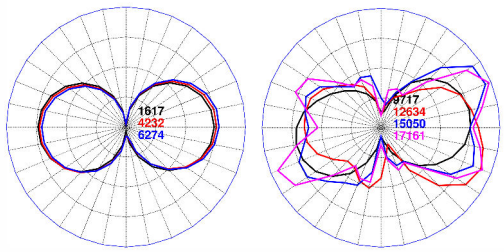


Figure 17. OctaSpHEAR v2 B format Y polar pattern

Version 2 shows slightly worse 1st order (Y) performance than version 1. The null of the lobes is shallower because the capsules have more omnidirectional behavior, and there is more spread in amplitude towards the back of the recovered lobes. On the other hand, version 2 shows better 2nd order (V) performance. The nulls are about 5dB deeper than in version 1, and the amplitude of the measured points is segregated into two groups only (ie: the shape of the lobes is more symmetrical than in version 1). The behavior near the 400Hz lower frequency limit is also marginally better.

doi:10.25836/sasp.2019.34

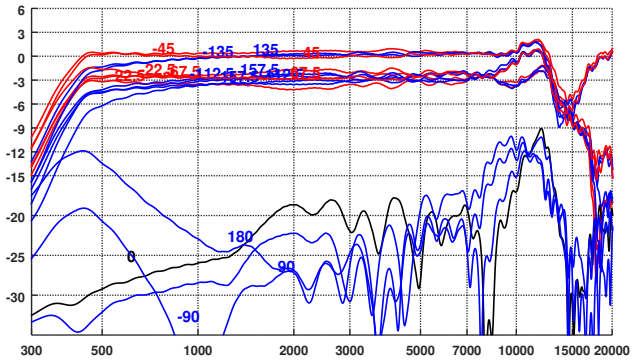


Figure 18. OctaSpHEAR v2 B format V frequency response with azimuth angle as a parameter

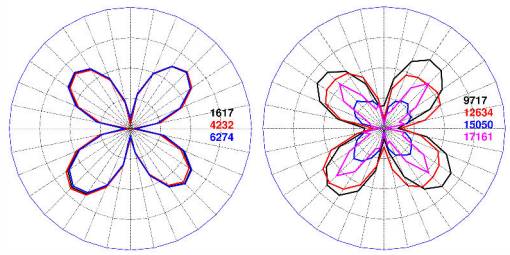


Figure 19. OctaSpHEAR v2 B format V polar pattern

It should be noted that the second order components only provide correct spatial information up to 11Khz, at which point we start seeing the effects of spatial aliasing.

The differences in the performance in the Ambisonics domain are smaller than expected, given the big differences in the behavior of the raw capsule signals, but the change in mechanical design in version 2 shows noticeable improvements, specially in the performance of the 2nd order signals.

4.3 Version 2, 3D performance plots

Figure 20 shows the 3D shape of the X 1st order component at low frequencies (from 800 to 1600Hz). The blue and green dots and their tessellation show the measured points, the red dots show the theoretical points of X for the full 240 point spherical t-design, normalized to the maximum of the measured points. The missing measurements (red dots without blue counterparts) are due to the limited reach of the robotic arm. Figure 21 and 22 show the measured and theoretical V and T second order components respectively. In all three plots there is a very good match between the theoretical and measured shapes.

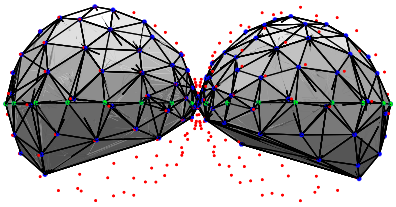


Figure 20. OctaSpHEAR v2 X component, measured (blue dots), measured in horizontal plane (green dots) and t-design (red dots)

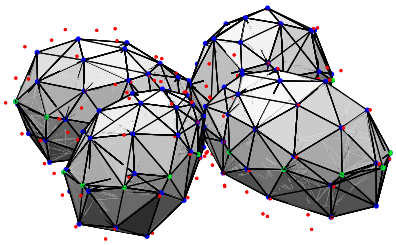


Figure 21. OctaSpHEAR v2 V component, measured (blue and green dots) and t-design (red dots)

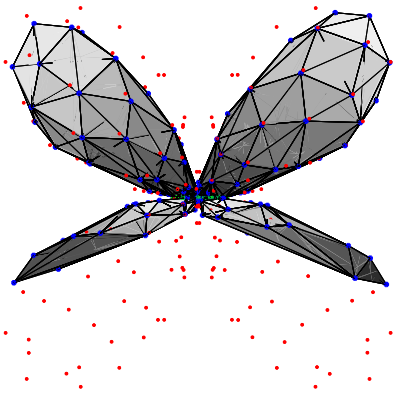


Figure 22. OctaSpHEAR v2 T component, measured (blue and green dots) and t-design (red dots)

4.4 Refining the encoder at very high frequencies

Close examination of the very high frequency behavior of Y shown in figures 12 and 16 indicates unexpected variations in the polar pattern.

Our software defines the shape of the B format equalization filters by measuring the power in logarithmically spaced bands for measurements near the peak of each recovered lobe. The inverse of this power profile is used to create the FIR filters. The following plots of Y versus frequency (from 5KHz to 20KHz) in all measured angles for elevations between -10 and 10 degrees (red traces), and between 20 and 40 degrees (blue traces), show how this approach fails at frequencies above 10KHz.

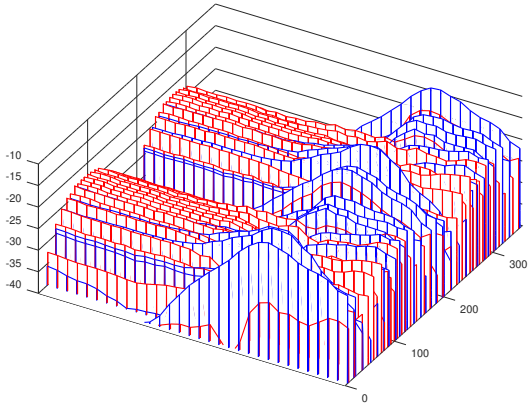


Figure 23. OctaSpHEAR v2 Y waterfall; red traces: -10 to 10 degrees elevation, blue traces: 20 to 40 degrees, 5KHz to 20KHz

doi:10.25836/sasp.2019.34

At the top of the spectrum we see substantial energy peaks outside of the two Y lobes, while at or near the peak of the lobes we sometimes have a drop in level. Our averaging algorithm does not take into account the energy outside of the lobes, and as a result that frequency range is boosted by the B format equalization filters.

The effect is much worse for elevations above (and below) the horizontal plane (blue traces). This is most likely the result of the on-axis resonant peak of the capsules at 12KHz.

The amount of unintended boost at high frequencies depends on the elevation angle, and our B format filters are angle-invariant, so it is impossible to compensate for this effect. However, we can design additional filters (and merge them into the corresponding B format filters) that take into account the peaks in that frequency range. Figure 24 shows the filter shapes we arrive at if we average all measurements for Y, V and W.

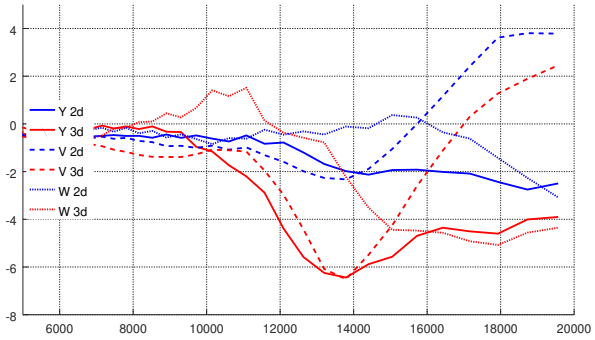


Figure 24. OctaSpHEAR v2 B format high frequency correction filters for Y, V and W

Red traces correspond to the full 3D measurement set, blue traces to the horizontal plane measurements. For both Y and V (and other components not shown) the correction filters for both sets of measurements agree on the frequency of the boost, but not for W (which should not be corrected). The amount of correction to be applied is a tradeoff between behavior in the horizontal plane and above and below it. Figure 25 shows the corrected spectrum if we choose a filter based on the full 3D set.

This is a calibration trade-off similar to the one that happens for the horizontal plane calibration of first order microphones. It is impossible to equalize equally well in all directions, and the choice of which signals to use to design the filters in that region will affect the behavior of the microphone.

5. CONCLUSIONS

Of the two designs we tested for this paper, version 2 (individual conical capsule holders with a central spherical core) has the best overall performance, specially at mid and high frequencies. We should stress that both versions (designs 1 and 2) perform very well, and the differences, advantages and disadvantages are subtle when listening to the results of a fully calibrated microphone. In particular, the first design (version 1) is surprisingly good in subjective terms, given the handicap of the polar response degradation at mid and high frequencies due to the resonances inherent in the mechanical design. We have also found additional corrections for the encoder that try to minimize unwanted effects at very high frequencies due to the deterioration of the capsule polar patterns at those frequencies.

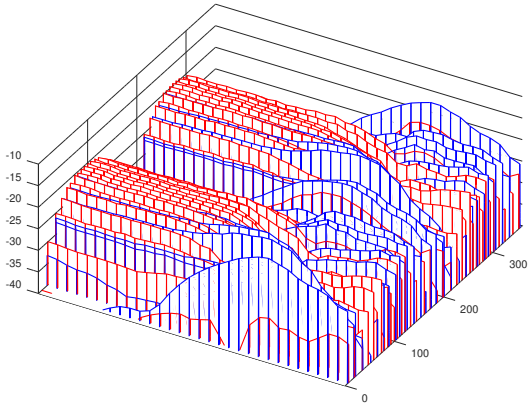


Figure 25. OctaSpHEAR v2 corrected Y waterfall; red traces: -10 to 10 degrees elevation, blue traces: 20 to 40 degrees, 5KHz to 20KHz

6. REFERENCES

[1] Michael Gerzon, “The Design of Precisely Coincident Microphone Arrays for Stereo and Surround Sound”, 50th Audio Engineering Society Convention, Preprint L-20, London, 1975

[2] Eric Benjamin, “A second-order soundfield microphone with improved polar pattern shape”, Audio Engineering Society Convention Paper 8728, 133rd Convention, San Francisco, 2012

[3] Fernando Lopez-Lezcano, “The *SpHEAR project, a family of parametric 3D printed soundfield microphone arrays”, AES Conference on Sound Field Control, July 18–20 2016, Guildford, UK

[4] Fernando Lopez-Lezcano, “The *SpHEAR project update: the TinySpHEAR and Octathingy soundfield microphones”, AES Conference on Audio for Virtual and Augmented Reality, August 20–22 2018, Redmond, Washington, USA

[5] Eric Benjamin, “Extending Quasi-Anechoic Electroacoustic Measurements to Low Frequencies”, Audio Engineering Society Convention Paper 6128, 117 Convention, San Francisco, 2008

[6] R. H. Hardin and N. J. A. Sloane, “McLaren’s Improved Snub Cube and Other New Spherical Designs in Three Dimensions”, Discrete and Computational Geometry, 15 (1996), pp. 429-441

[7] Aaron Heller, “Derivation of the A-to-B matrix for a coincident array of first-order microphones”, unpublished, 2007

doi:10.25836/sasp.2019.34

SPACES @ CCRMA: DESIGN AND EVOLUTION OF OUR 3D STUDIO AND CONCERT DIFFUSION SYSTEMS

Fernando Lopez-Lezcano
CCRMA, Stanford University
nando@ccrma.stanford.edu

ABSTRACT

This paper describes the design and evolution of three multi-speaker systems that provide 3d surround sound diffusion capabilities and are used for research, composition and concert diffusion at CCRMA, Stanford University. The first of the two permanently installed systems is housed in the Listening Room, a full 3d studio with 22 speakers and 8 subwoofers, including 7 speakers mounted below the acoustically transparent floor. The second permanent system is in our small in-house concert hall, the Stage, which after a recent full audio remodel is equipped with 56 speakers and 8 subwoofers. Finally, our concert diffusion system, the GRAIL (the Giant Radial Array for Immersive Listening), is a "portable" system that can be set up and calibrated in a couple of days and can currently support up to 32 speakers and 8 subwoofers.

1. INTRODUCTION

All three systems have evolved separately but share a common philosophy with regards to sound diffusion tasks. In most conventional studios the computers where composers, sound artists, or mixing engineers work, are either directly connected to the speakers, or are routed to them through a conventional digital mixer. In the systems described in this paper there is a diffusion layer with its own computer and custom software that sits in between the computers where work is done, and the speakers themselves, and there is no digital mixer. The user interface is very minimal and simple, and motorized fader boxes provide real-time control interaction when it is required.

The diffusion computer is tasked with delivering the best possible sound quality in each space. All speakers are delay and level matched, and digitally equalized to be as neutral as possible. The software handles digital phase matched crossovers to the subwoofers, and provides calibrated and tuned Ambisonics decoders of different orders and configurations, ready to use. The system is a black box that accepts multiple digital multi-channel streams and renders them transparently to the available speakers. It does not have provisions for doing spatialization tasks, and does not include the level of sound processing present in a digital mixer, such as equalization and compression, as those functions are best performed in the work computers.

The addition of Ambisonics modes with calibrated decoders for each space make it particularly easy to move pieces between them and fully use the capabilities that each spaces provides. Pieces can be composed and auditioned in Ambisonics in the Listening Room or the Stage, and then performed in concert with the GRAIL system with minimal or no changes.

In this paper we will describe the evolution of the three systems, and the current state of hardware and software that supports composition, experimentation, research and sound diffusion tasks at CCRMA.

2. THE LISTENING ROOM

The Listening Room was built when our building, The Knoll, was remodeled in 2005. It was designed to be an acoustically dry, low noise floor, high quality full 3D studio. Part of the floor was opened and covered with an acoustically transparent grid, so that speakers could be installed below the listening position. 16 speakers were initially installed in 2006, 4 below the grid floor, 8 around the listening position and 4 more hanging from the ceiling, all of them driven by a 16 bus Tascam DM3200 digital mixer.

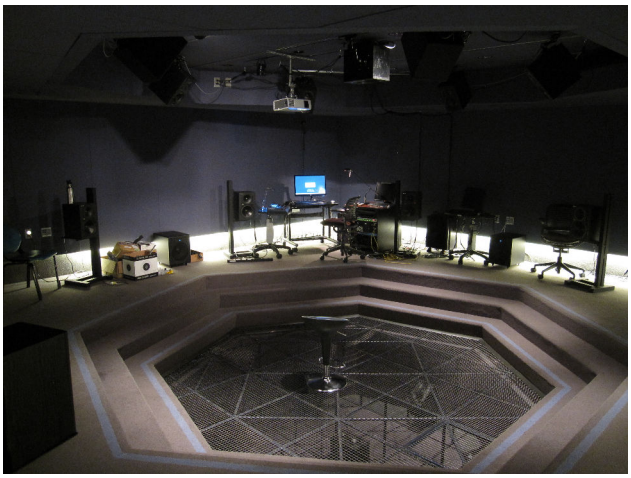


Figure 1. The Listening Room

After the system had been in use for a while it became increasingly clear that a digital mixer was not the best option for controlling the system. It provided functionality

© Fernando Lopez-Lezcano. Licensed under a Creative Commons Attribution 4.0 International License (CC BY 4.0). **Attribution:** Fernando Lopez-Lezcano. "Spaces @ CCRMA: design and evolution of our 3d studio and concert diffusion systems", 1st EAA Spatial Audio Signal Processing Symposium, Paris, France, 2019.

doi:10.25836/sasp.2019.29

that was not needed, it was unnecessarily complicated to operate and was not expandable. In 2008 we started to design and implement a fully custom sound diffusion system based on one of our fanless Linux workstations. The goal was to have a system that was easy to use and provided a very simple interface to the task of diffusing sound in the studio.

2.1 The OpenMixer hardware and software

The initial design [1] used a computer as the central hub to the diffusion tasks, with a simple user interface provided by two USB fader boxes. We selected SuperCollider as the computer language to use for its implementation, and also tried to leverage as much free software as possible to simplify the design task (Jconvolver, Ambdec and others). Everything was connected together through Jack, a low latency sound server, and the system was designed to boot unassisted into the control software.

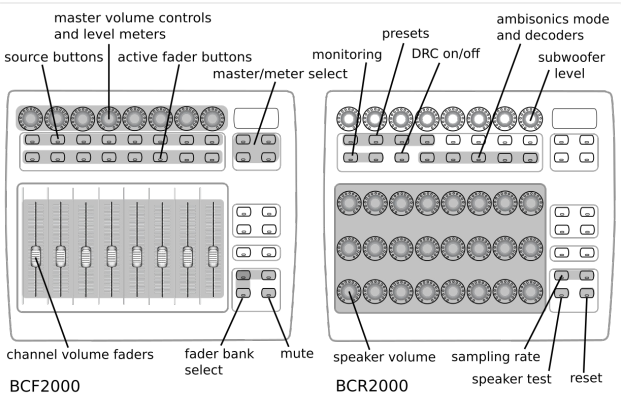


Figure 2. Listening Room: OpenMixer user interface

Two PCI RME audio interfaces with MADI and ADAT I/O, and SSL converters, provided the audio I/O. Audio could be connected to the system through 16 analog inputs, two ADAT digital interfaces with word clock or an ethernet connection using Netjack Jack clients - the easiest way at the time to connect to the system 24 channel audio coming from external laptops and computers. A dedicated 24 channel ADAT connection was provided for a desktop workstation permanently installed in the studio, so that users would be able to use the system with minimal training.

The system was designed with two modes of operation. The Direct mode allows any input to be routed directly to any speaker, with arbitrary mappings. The Ambisonics mode routes the inputs to an Ambisonics decoder tuned to the speaker array. This makes it easy to work in Ambisonics without having to deal with designing and tuning decoders for the speaker array.

The system proved to be stable, and it was expanded in 2011 [2] to 22 speakers and 4 subwoofers. The spatial arrangement of the speakers was chosen to be 1 + 6 + 8 + 6 + 1, selected so it could provide accurate decoding of 3rd order periphonic Ambisonics audio streams.

In 2014 digital speaker equalization was added to the system. In addition to matching the delays and levels for all speakers, digital equalization filters were designed for each speaker using the DRC (Digital Room Correction) software, to provide the best frequency and phase response possible. The sound quality improved significantly, specially for the speakers installed below the floor which had always been problematic due to the relatively small size of the pit in which they were housed and the coloration it created.

3. CONCERT SYSTEMS



Figure 3. D.C Power Lab, preparing for a concert, circa 1980

Through CCRMA's history immersive music production and subsequent concerts have been part of daily life. Concerts were staged in the D.C Power Lab and later in the Frost open air auditorium using a four channel system, and also in Dinkelspiel and Campbel Recital Hall in the Music Department. The Ballroom, a small space that was used for teaching and concerts in our building, and later our backyard at The Knoll, hosted many events that used arrays of up to 8 speakers driven from analog and later digital mixers.

The planned construction of a new concert hall at Stanford, the Bing Concert Hall, spurred us to start planning events for the inaugural season which was to happen in 2013. Starting in 2009 we bought 8 new speakers for our concert system, then expanded to 16 and finally to 24 (with 8 matching subwoofers), which would enable us to diffuse sound in full 3D surround in the Bing spaces. Our Transitions outdoor concerts in 2011 and 2012 were testing grounds for new diffusion technologies, including higher order Ambisonics rendering and digital real-time simulation of real acoustic spaces. We started writing software inspired in our Listening Room control system to use one of our high performance custom fanless workstations [3] as the diffusion engine for concerts, including software that allowed us to use a 32 channel digital snake system as a low cost high channel count D/A and A/D [4]. In 2011 we successfully used this new system to diffuse a whole concert, including completely remoting the computer as all peripherals could interface with it through ethernet cables.

In 2013 we staged several events in the Bing Main Hall and its black box rehearsal space, the Studio. In both cases

doi:10.25836/sasp.2019.29

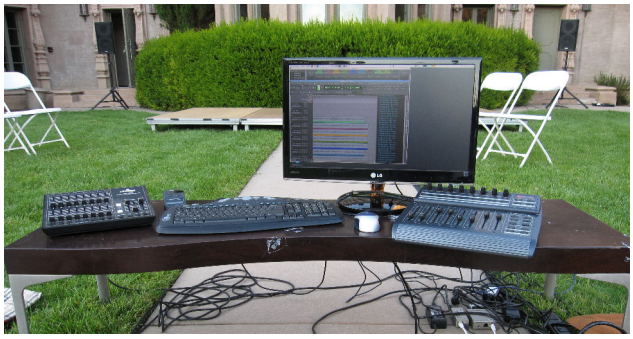


Figure 4. Transitions 2011 concert diffusion control system

with dome speaker configurations of 12 + 8 + 4 + 1 which enabled us to render pieces with the same type of flexibility we enjoyed in our Listening Room, but for a large audience. Probably the most ambitious was the digital simulation of the acoustic space of Hagia Sophia [5] (part of the Icons of Sound project), in which the Cappella Romana group performed in real-time, singing byzantine chanting that was composed ages ago for that space. Their voices were input into our system which was mixing and processing multiple impulse responses, finally routing them to a 24 speaker dome, in effect creating a virtual space for the singers to perform in.



Figure 5. Bing Main Hall, "Bada Boom Bada Bing!" concert, 2013

The diffusion system, later called the GRAIL (the Giant Radial Array for Immersive Listening) evolved over the years and currently can be deployed with up to 32 speakers and 8 subwoofers [6]. It is calibrated digitally after it is installed, with the goal of bringing high quality studio sound to the concert hall. In addition to the regular concerts at Stanford, we have staged events abroad, first in the context of the 2016 Sound Symposium in Newfoundland, and more recently in Taipei, Taiwan, as part of the 2018 season Innovation Series events in the National Theater and Concert Hall (NTCH).



Figure 6. Bing Studio, "CCRMA Spring Concert", 2013

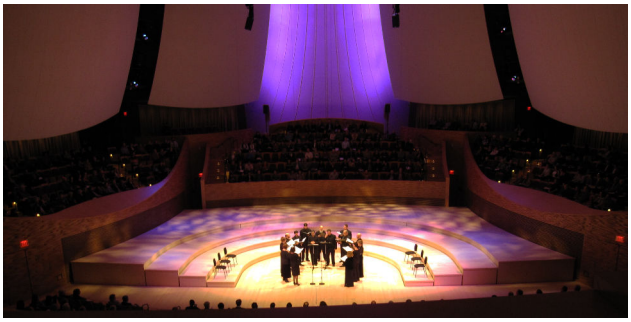


Figure 7. Bing Main Hall, "From Constantinople to California", Cappella Romana, 2013

4. THE STAGE

The Stage, our small in-house concert hall, also built during the 2005 renovation, is a cathedral ceiling room with very good acoustics and a low noise floor that can seat between 40 and 80 people depending on the type of event being staged. It is a multi-purpose space that can be used for concerts, lectures and classes. Its initial sound diffusion system was the same 8 channel ring of high quality Adam mid-field studio monitors that used to be housed in our Ballroom before the remodel.

The system was upgraded in 2009 to 16 speakers in a two ring configuration (8 + 8), with the upper speakers hanging from the ceiling pipes and with 8 subwoofers added to the low ring. The upper speakers were later repositioned for better rendering of 3d spatial sound, eventually hanging in an 8 + 6 + 2 configuration. A Yamaha DM1000 digital mixer is used to diffuse sound and is directly connected to the speakers.

4.1 The GRAIL comes to the Stage

While we had been using the Listening Room for many years, and the new GRAIL was being deployed for concerts several times every year, our diffusion systems had some shortcomings. In terms of spatial resolution both systems could accurately decode Ambisonics up to third order (and fourth order if some errors were ignored), but

doi:10.25836/sasp.2019.29



Figure 8. The Stage in 2011, 16 speakers in 8 + 8 configuration, 8 subwoofers in the bottom of the main towers

composers and sound artists had started using higher orders routinely. We could not play their pieces in their original resolution, or compose HOA content with our existing facilities. We were also limited in physical size. Our only 3D capable permanent setup was the Listening Room, and the bigger concert spaces were only available for a few days before the concerts. Hardly what is needed to compose and audition new music.

At the end of 2015 we proposed to upgrade the diffusion capabilities of the Stage to convert it into a state-of-the-art 3D space which could be used for concerts, research and composition. Adding 32 small speakers to the existing system would transform it into a 48.8 system that would allow it to render 5th or 6th order Ambisonics pieces correctly, or the equivalent in spatial resolution using other systems [7].

But the Stage is not only a concert hall, it is also regularly used for classes, lectures, demos and other events that do not need or want a high spatial resolution speaker array. In fact, the majority of users require access to just stereo playback. As the CCRMA concert events combine live performers, touring musicians and researchers, many concerts do not deal with 3D surround sound, and use mostly stereo projection.

It quickly became apparent that the existing system could not be replaced. In fact a key requirement for the project to go forward was to leave the existing system in place, with no user visible changes. And while it was obvious that the full system would require a control computer and software similar to the one running in the Listening Room, the legacy system had to work even if the control computer was turned off.

All those requirements together made for quite a challenging project.

One of the key tasks of the design process was finding an audio transport technology that could meet these requirements at a reasonable cost. After much research we selected AVB (Audio Video Bridging) [8] [9] as a suitable technology that could interface with our open Linux-based systems, specially since there was a family of products manufactured by Motu that seemed to meet our requirements.

We created our audio system by interconnecting several Motu audio interfaces through ethernet, and using their USB interfaces as entry points for computers into the system. The internal routing matrices in the audio interfaces were used to reconfigure connections and switch operating modes. And since the Motu cards can be remotely controlled, a small Raspberry Pi powered computer with a touch screen could be used to switch operating modes.

Figure 9 shows the hardware of the minimal system in "Digital Mixer Mode", in which the old legacy 16.8 system is controlled through the DM1000 mixer.

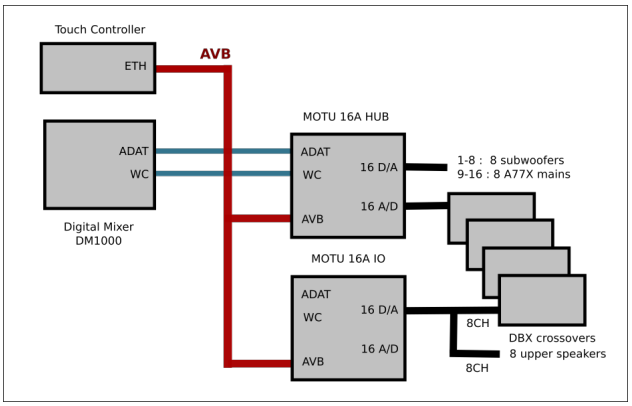


Figure 9. Stage audio system: signal routing in Digital Mixer mode

Figure 10 shows the expanded system in OpenMixer mode. The audio interfaces used for the legacy mode are now a subset of the full system and their audio routing is completely different. All signals are now processed by the control computer before being sent to the speakers.

The upgrade was done in several stages, it took almost two years, and there is still software work to do. During this time we also upgraded nearly all of the original speakers for a much better basic sound quality.

The current Stage audio system consists of 8 Motu audio interfaces interconnected through AVB and ethernet cabling, 56 main speakers of different sizes, and 8 subwoofers. Control and user computers are connected to the system through the USB interfaces of the Motu boxes, which can support up to 64 channels under Linux with the proper firmware. All speakers are digitally equalized and we can easily run 5th or 6th order Ambisonics decoders tuned to the system.

The inaugural concerts happened at the end of 2017, after the first stage of the upgrade, and the system has been operating successfully since then.

5. EXPANDING THE GRAIL

The experience we gained with AVB systems during the Stage upgrade has percolated to our other diffusion systems as well.

The GRAIL has moved away from our custom software and digital snake solution, which proved to be very reliable for years but limited us to 32 channels of I/O, to a

doi:10.25836/sasp.2019.29

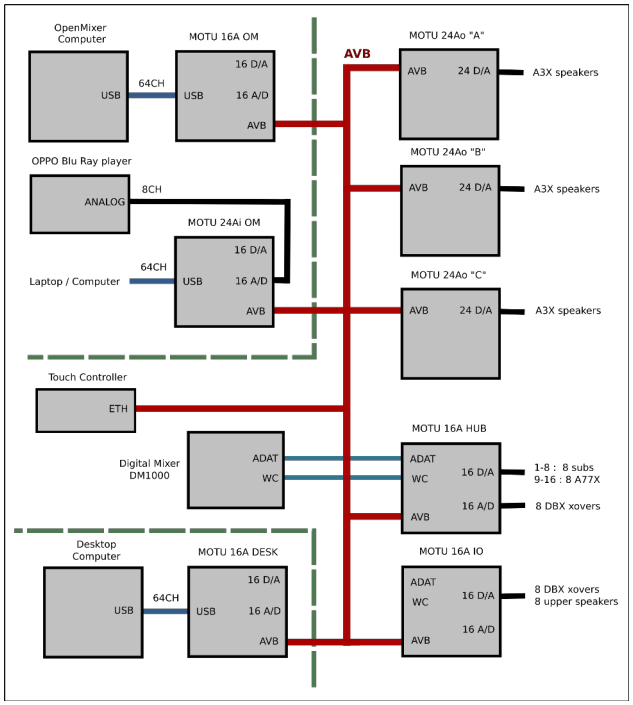


Figure 10. Stage audio system: signal routing in OpenMixer mode



Figure 11. the Stage, "Transitions" concert, 2018

small network of between 4 and 6 Motu audio interfaces interconnected through ethernet cables. This has proved to be very flexible as it enables us to position the audio I/O where it is needed, makes all the analog connections shorter and enables us to connect multiple computers to the core system and route them appropriately as needed using AVB streams and the internal connection matrices of the interfaces. It also allows us to expand beyond the previous limit of 32 channels of I/O.

The control computer has changed as well. We had been using one of our high powered Linux fanless workstations as the control computer for both tasks: GRAIL core functionality and the diffusion of pieces in concert (this is a 4 or 6 core computer with 32 or 64G of RAM and fast SSD disks). This reduced the reliability of the core system as the computer had to deal with very dissimilar tasks that can interfere with each other.

We are now using two computers instead. The much smaller, more portable machines are based on Intel quad core NUC systems, and are housed in special purpose passively cooled fanless cases (the original NUCs are very small but use a noisy fan to cool the processor). The main control computer runs the core software (written in SuperCollider) which handles audio routing, master gain control, delay and level matching, speaker digital equalization, digital crossovers, and Ambisonics decoding. A second NUC system is used to diffuse pieces, either by playing back fixed media content through software such as Ardour, or acting as a digital mixer for live pieces, or combinations of the above. Both systems run Linux using real-time optimized kernels, and with proper tuning of all system priorities so that we can run at very low latencies.

Both computers connect to separate Motu audio interfaces, and their audio streams are routed through AVB and ethernet cabling. In fact there is nothing special about the diffusion computer, and with AVB we can have several computers, including performers' laptops, connected to the system, and seamlessly switch between them.



Figure 12. GRAIL concert diffusion control interfaces and computers

We have used this system in our last round of concerts and it has proven to be very configurable and reliable. We have also recently added a small single fader control surface to the control computer, which we configured as a master level controller. We can also select the operating mode, type and order of the Ambisonics decoder, and we are going to add presets for diffusion for common speaker configurations, such as a ring of 8 speakers or 5.1/7.1. This makes it much easier to control the system during a concert.

6. UPGRADING THE LISTENING ROOM

The experience of using the new Stage system led us to upgrade the system in the Listening Room as well. One feature shared by the new Stage system and our updated GRAIL was the ability to connect multiple external computers to the system through USB audio interfaces. That is a much easier path to connecting to the system than the old NetJack ethernet interface in the Listening Room.

doi:10.25836/sasp.2019.29

We upgraded the whole system, adding a new, faster rack-mounted control computer and replacing the audio hardware with Motu interfaces. The OpenMixer control software was upgraded to use the new hardware and moved to the new computer. The user interface did not change, and the only compromise was to only allow mixing of different sources up to a maximum of 64 channels (the old system could mix all sources at the same time).

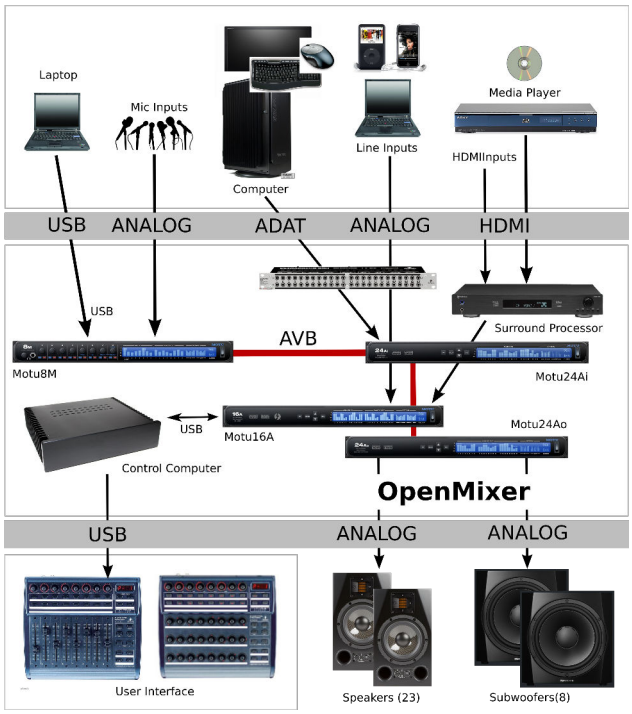


Figure 13. Listening Room hardware and audio routing block diagram

We also replaced all the speakers by newer, much better sound quality Adam A7X/A5X units, and replaced the original four subwoofers with eight new new Dynaudio 9S units.

7. CONCLUSION

The three systems provide different full 3D surround calibrated environments for composers, sound artists, and researchers to work in. The Listening Room is a very dry full 3D experimental space, the Stage provides a larger space with higher spatial resolution that can also host concerts, and the GRAIL can travel to even bigger spaces and render pieces composed in the other spaces without much change, allowing us to host concerts for much larger audiences. All three systems are in constant use, and in particular our new Stage array has been used frequently in concert over the past 1.5 years.

8. THANKS

Many people contributed over the years to make all of this happen. Jason Sadural kick-started the initial OpenMixer code in 2008, and Elliot Kermit-Canfield helped transform the Listening Room in the 2011 upgrade. Many thanks to Carr Wilkerson for his help in caring for the GRAIL for many years, to Eoin Callery for taking over after Carr left, and to Constantin Basica for doing the same after Eoin started teaching. Without Christopher Jette the upgrade to the Stage to a state-of-the-art system would not have happened at all, and we can touch control it thanks to the work of Carlos Sanchez Garcia-Saavedra. Matt Wright helped immensely with the 2018 round of upgrades to the Listening Room. Thanks also to the many students that have coiled and uncoiled miles of cables over the years, shouldered speakers, and soldered many small connectors during our upgrades. And most of all thanks to the countless composers that have contributed pieces over the years, making it worthwhile to spend so much time on this. At the end of the concerts we are all smiles.

9. REFERENCES

[1] F. Lopez-Lezcano and J. Sadural, "Openmixer: a routing mixer for multichannel studios," in *Proceedings of the Linux Audio Conference 2010*, 2010.

[2] E. Kermit-Canfield and F. Lopez-Lezcano, "An update on the development of Openmixer," *Proceedings of the Linux Audio Conference 2015*, 2015.

[3] F. Lopez-Lezcano, "The quest for noiseless computers," in *Proceedings of the Linux Audio Conference 2009*, 2009.

[4] F. Lopez-Lezcano, "From Jack to UDP packets to sound and back," in *Proceedings of the Linux Audio Conference 2012*, 2012.

[5] F. Lopez-Lezcano, T. Skare, M. J. Wilson, and J. S. Abel, "Byzantium in bing: Live virtual acoustics employing free software," in *Proceedings of the Linux Audio Conference 2013*, 2013.

[6] F. Lopez-Lezcano, "Searching for the GRAIL," *Computer Music Journal*, vol. 40, no. 4, pp. 91–103, 2016.

[7] F. Lopez-Lezcano and C. Jette, "Bringing the GRAIL to the CCRMA Stage," in *Proceedings of the Linux Audio Conference 2019*, 2019.

[8] "The AVNu Alliance (AVB)." `https://avnu.org/`.

[9] "OpenAvnu GIT repository." `https://github.com/AVnu/OpenAvnu`.

doi:10.25836/sasp.2019.29

A VERY SIMPLE WAY TO SIMULATE THE TIMBRE OF FLUTTER ECHOES IN SPATIAL AUDIO

Tor Halmrast
University of Oslo/Musicology
torhalm@online.no

ABSTRACT

The "strange" timbre of flutter echoes is often not included in spatial audio, auralisations and room simulations for video games etc., perhaps due to lack of knowledge, but also because a detailed simulation will be very heavy. In common room acoustic modelling only a relatively small number of reflections are used, and the later part of the decay is treated simply as diffuse reverberation. For rooms likely to give a flutter echo, this will not be sufficient. This paper will explain why a flutter echo gives the characteristic mid-/high frequency "tail" and show how this can be simulated adding a band pass filtering to a "ping-pong" echo between two loudspeakers.

1. INTRODUCTION

Flutter echoes are usually thought of as a defect one simply wants to avoid. The physics of flutter echoes is, however, not simple. Repetitive reflections with Δt [s] between each reflections give a perceived tone with a frequency of fo=1/Δt [Hz] and multiples of this. Often this "Repetition Pitch"/"Repetition Tonety[1]" is used to explain the "tonal" character of a flutter echo in rooms with two parallel, reflecting surfaces and the other surfaces almost totally absorbing. However, fo=1/Δt is in the low frequency range but the characteristic "almost tonal" character of a flutter echo is of mid/high frequency, typically around 1-2 kHz. Also sound engineers mix up these effects, and several plug-ins called "pong-echo" etc. forget this special timbre of real flutter echoes. This paper gives an overview on several ways to explain the special timbre of flutter echoes, by inspecting Diffraction, Mirror sources, Fresnel Zones, Transformation from spherical to plane waves etc. This knowledge about flutter was implemented as a sound effect not only in time but also frequency domain.

The paper shows measurements of flutter in actual rooms compared with simulations in room acoustics modelling software (Odeon), empirical evaluations, Fresnel-Kirchhoff approximations of diffraction and simulations in MatLab (Edge Diffraction Toolbox). Each of these methods does not fully describe the physics of flutter, but together they give interesting views on what is happening. For a deeper analysis, see [1] and [2]. This paper shows that the resulting characteristic mid/high frequency timbre of a flutter in ordinary rooms is not a "tone", but a gradual band pass filtering of the broad banded impulsive signal, like a gradual subtractive synthesis. We find that this filtering is a combination of two filtering effects: Low frequency dampening due to the increasing source distance and diffraction, which gives that the sound field is transferred from spherical to plane waves, and High frequency dampening due to air absorption, as shown in the overview in figure 1.

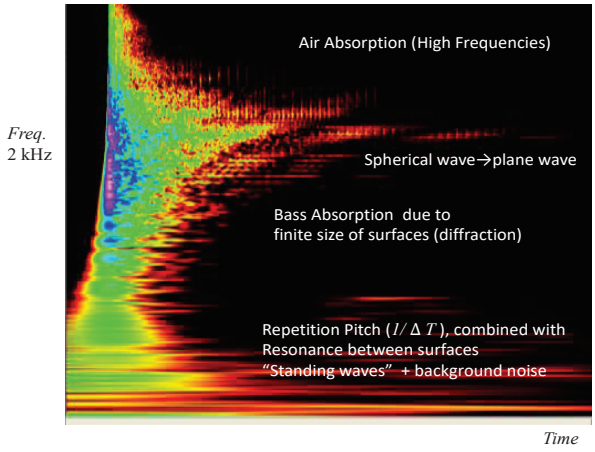


Figure 1. Overview of the timbre of flutter echoes

The sound pressure level of a plane wave is reduced only by air absorption and the absorption at the surfaces, while a spherical wave is reduced by 6 dB per doubling of distance. Together these two main filtering effects give the characteristic mid/high frequency "almost tonal" character of flutter, which we will call the "Flutter Band Tonety[1]" (or just "Flutter Tonety"), as a distinction from the "Repetition Tonety". Depending on the amount of bass in the signal, its duration and especially the position of the sender/receiver with respect to the resonance peaks and nodes of the standing wave pattern of the room resonances between the surfaces (and "overtones" thereof) will appear, but for most positions between the reflecting surfaces, and especially for short sounds like handclaps, the "Flutter Band Tonety-tail" in mid/high frequencies will last longer.

© 2019 Tor Halmrast. Licensed under a Creative Commons Attribution 4.0 International License (CC BY 4.0). Attribution: Tor Halmrast. "A very simple Way to Simulate the Timbre of Flutter Echoes in Spatial Audio", 1st EAA Spatial Audio Signal Processing Symposium, Paris, France, 2019.

[1] The word "Tonety" is chosen by the author for such "almost a tone" because Pure Tones, Pitch and Tonality have more precise definitions.

doi:10.25836/sasp.2019.37

2. MEASUREMENTS OF FLUTTER ECHOES

Measurements of several flutter echoes in real rooms and in anechoic chamber are given in [1]. A typical measurement of a flutter echo in a foyer with absorbent ceiling and two reflecting, parallel walls is shown in fig. 2.

We see that the decay ends up in a "tail" around 2 kHz, almost like a gradual subtractive synthesis. If the surfaces are somewhat absorbing for high frequencies, this "tail" appears at a somewhat lower frequency.

(PS! The distance between the walls was app. 12 m, giving a room resonance between the two surfaces of app. 14 Hz, which proves that the room resonance/Repetition "Tonety" is far away from the flutter tail of 2 kHz).

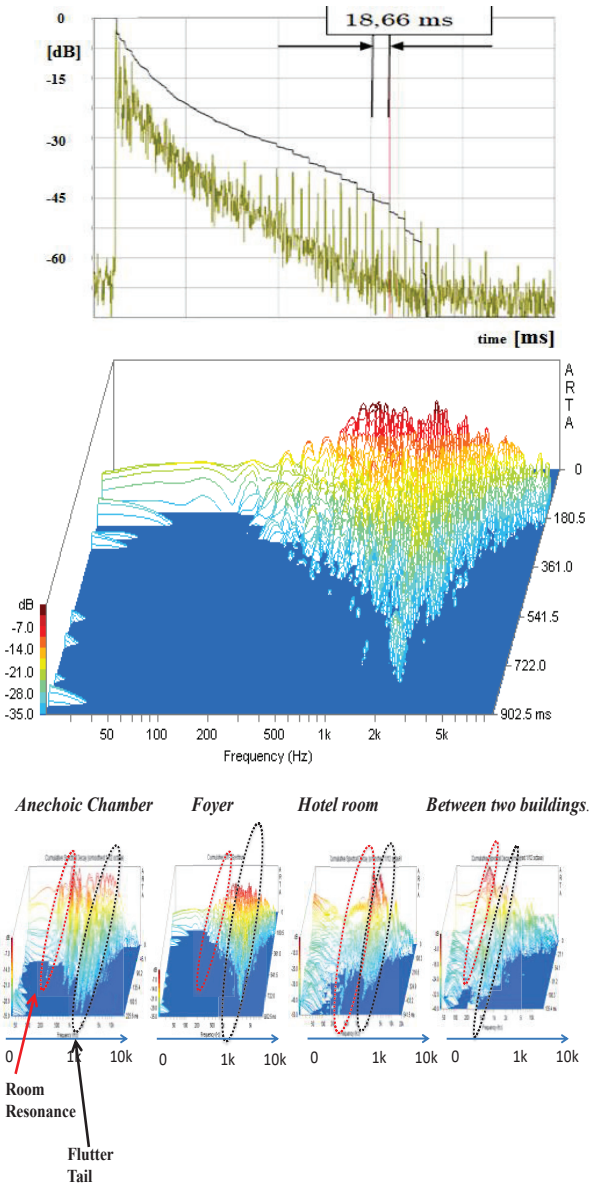


Figure 2. Upper panes: Measurement of typical flutter echo in a room: Impulse Response, and Waterfall. Lower pane: Waterfall curves of flutter echoes in several rooms.

3. SPERICAL TO PLANE WAVE

A very simple Odeon [3] room acoustics model with two parallel, reflecting surfaces was prepared (all other surfaces totally absorbing). Figure 3 shows the radiation from a point source (spherical wave). The receiver position is almost the same as the sender (as for a person clapping) and these are both positioned closer to the bottom of the surfaces, giving the possibility to inspect the situation both for a small surface (in the upper part of each figure) and a bigger surface (in the lower part of each figure).

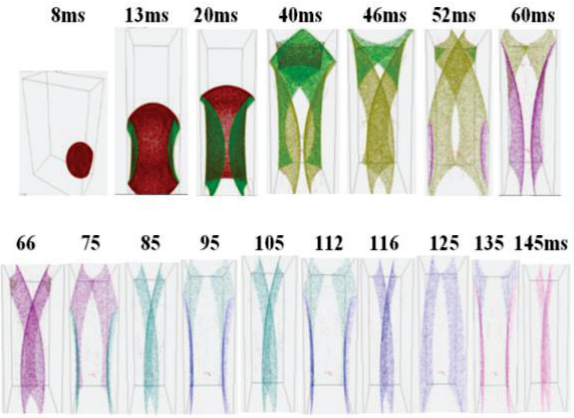


Figure 3. Odeon simulation of flutter between two parallel, reflecting surfaces, showing the transformation from spherical wav to plane wave.

Figure 4 shows similar Odeon simulation, where we see the diffraction from the edges more clearly (the small. "later" reflections, gradually decreasing).

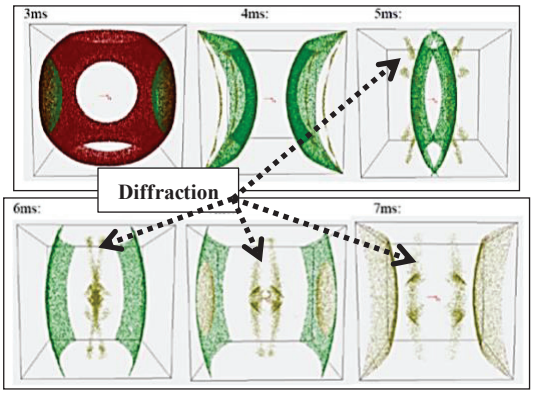


Figure 4. Odeon simulation showing the diffraction from the edges of the reflecting surfaces, which gradually combine in a destructive way, leaving the plane wave in the last part of fig.3.

4. SPERICAL TO PLANE WAVE

When you clap your hands at a distance from a surface, the result will be the combination not only of the direct sound and the reflected sound, but also the diffraction; which is the "reflection" from the edge of the surface. For some frequencies they arrive in phase and for other frequencies, one or two of them might arrive out-of-phase with another.

doi:10.25836/sasp.2019.37

A typical situation for one reflecting surface is shown in figure 5.

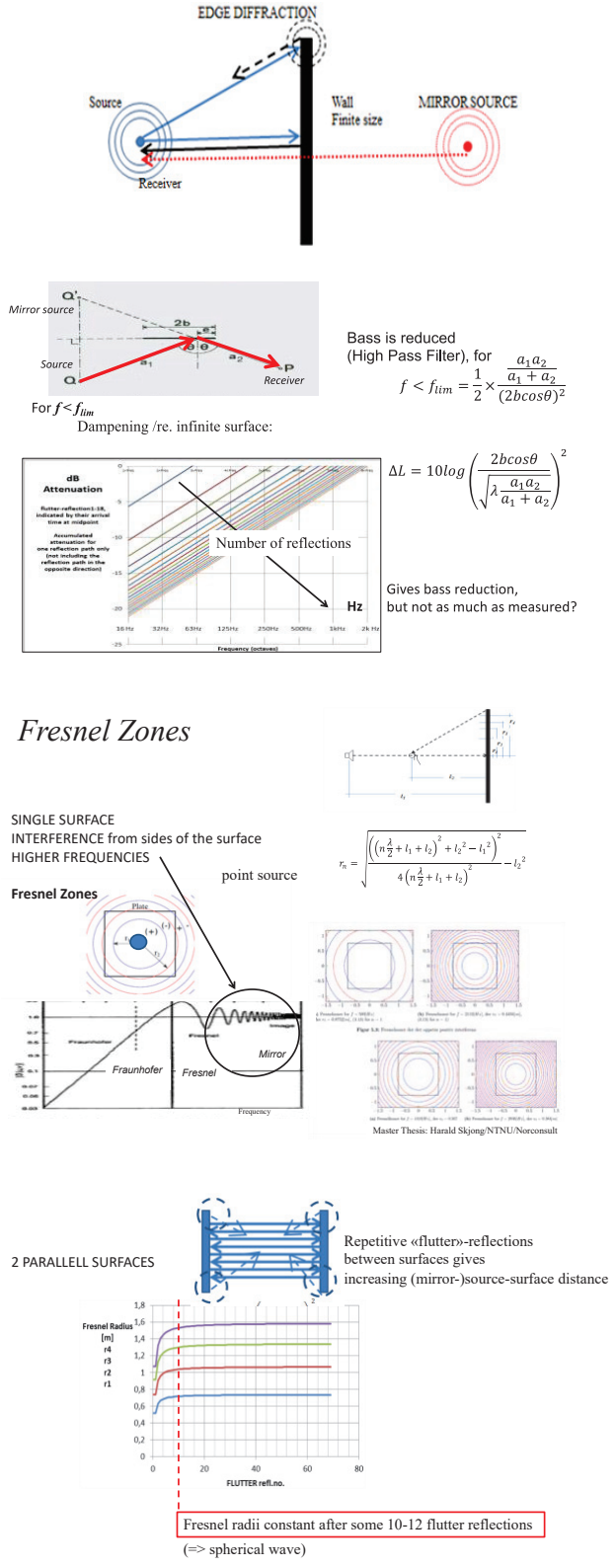


Figure 5. Mirror source and Diffraction from the edge of a finite surface. Receiver is at source position (as for a person listening to his own handclap). Middle and lower panes: Illustrations of diffraction. See [1] for discussions.

Fig. 4 showed how diffraction influenced the fluttering reflections between two walls. For our repetitive reflections the distance between mirror source(s) and its corresponding reflecting wall grows very rapidly, giving that, seen from the mirror source, the surface(s) appears smaller and smaller. The result is that they reflect gradually less and less in the bass and lower mid-frequencies. Several methods for calculations of diffraction are given in [1], and are beyond the scope of this paper. The results [1] from the analysis in MatLab (Edge Diffraction Toolbox by Peter Svensson [7]), iterative use of Rindel´s approximations of Fresnel/Kirchoff [8] etc. confirm the main conclusions about flutter echoes.

5. INFLUENCE OF DIMENSIONS AND ABSORPTION; KUHL´S EQUATION

Flutter was investigated by Maa [4], Krait et al. [5] and Kohl [6]. Both [5] and [6] states that for a plane wave between two surfaces ***S*** [m2] with distance ***l*** [m], the wave is dampened only by the absorption coefficients ***α*** at each surface and the air absorption, ***m***. (frequency dependent; ***4m*** is typically 0 for low frequencies, 0.01 for 1 kHz, rising to 0.03 for 4 kHz). More background for Kuhl´s equations is given in [1]. The frequency content of flutter can be looked upon as the "sum" of three reverberation "asymptotes" for the reverberation time versus frequency, ***f***.

1. Low Frequency damping due to finite surface area:

$$T_1 = \frac{0.041\ x\ 2fS}{c} \quad (1)$$

(where c is the velocity of sound, typically 343 m/s)

2. Damping due to absorption (α) on the surfaces:

$$T_2 = \frac{0.041 \times l}{\alpha} \quad (2)$$

3. Damping in the air (dissipation):

$$T_3 = \frac{0.041}{m} \quad (3)$$

The total reverberation time, T_{FL}, can be re-written *as*:

$$\frac{1}{T_{FL}} = \frac{1}{T_1} + \frac{1}{T_2} + \frac{1}{T_3} \quad (4)$$

Fig. 6 shows how these three "asymptotes" work together to get the total maximum reverberation for a mid/high frequency band, and how the different parameters influence on the position of the "peak" and, to a certain degree, how narrow this "tail" will be, (the "Q-factor" of the combined filter) (logarithmic freq.- and T-axis).

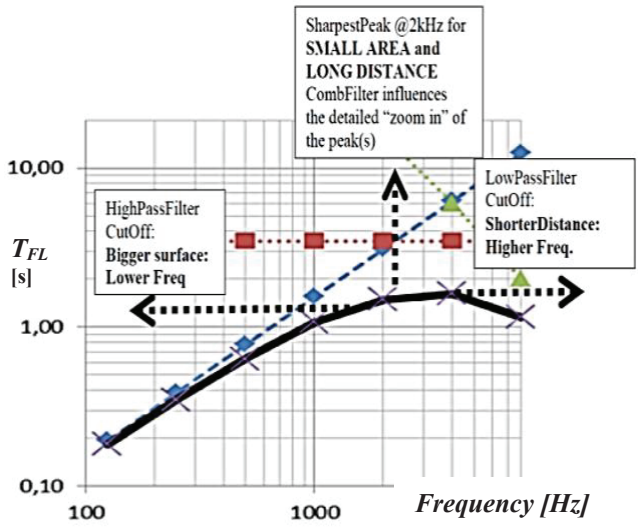


Figure 6. Illustration of Kuhl´s equation, showing how the different parameters influence the reverberation time of flutter echoes

6. LINKS BETWEEN ROOM RESONANCES AND FLUTTER TAIL

The waterfall curves in fig. 7 show the two main "tonalities" a flutter echo. The lowest "hill" (marked 1, red/dotted ellipse) indicates the "Repetition Tonety" (*fo*=1/Δt) between the surfaces. (Often disturbed by some background noise). For gradually higher frequencies we see the "harmonics" of this resonance (2*fo*, 3*fo* etc.) We see that the mid/high band (marked 2, black/solid line ellipses) last longer and one of these "overtones" will of course "win" in the competition of lasting the longest. The fact that a mid/high frequency band last longer than the fundamental of the Repetition Pitch/"Tonety" (*fo*), is therefore not a direct result of the *fo*-resonance itself, but as we only have the multiples of *fo* to choose from towards the "tail", there is of course a certain link between the two main "tonalities" of flutter. It is like a subtractive synthesis gradually resulting in one (or some) of the many higher overtones of the room resonances.

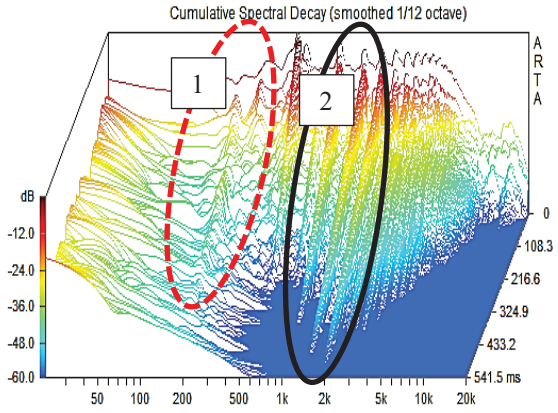


Figure 7. Waterfall curves of flutter "Resonance Tonety" (1, red/dotted), "Flutter Band Tonety" (2, black/solid line).

Fig. 8 shows an overview of these two main "tonalities" of flutter, now using a linear frequency axis. The equally spaced lines are the "overtones" of the "Resonance Tonety" *fo*. The overall filtering giving the mid/high frequency "tail" is the "Flutter-Band-Tonety" as a result of the High Pass Filter due to non-infinite surfaces and increasing distance between mirror source and surface for each flutter reflection, and the Low Pass filtering due to air absorption. The impact of the "Repetitive Tonety" (marked 1), combined with the room resonances is highly dependent on the signal and the positions of sender and receiver, but the flutter filtering towards the "tail"/"Flutter Tonety" (marked 2) is perceived much easier for all positions.

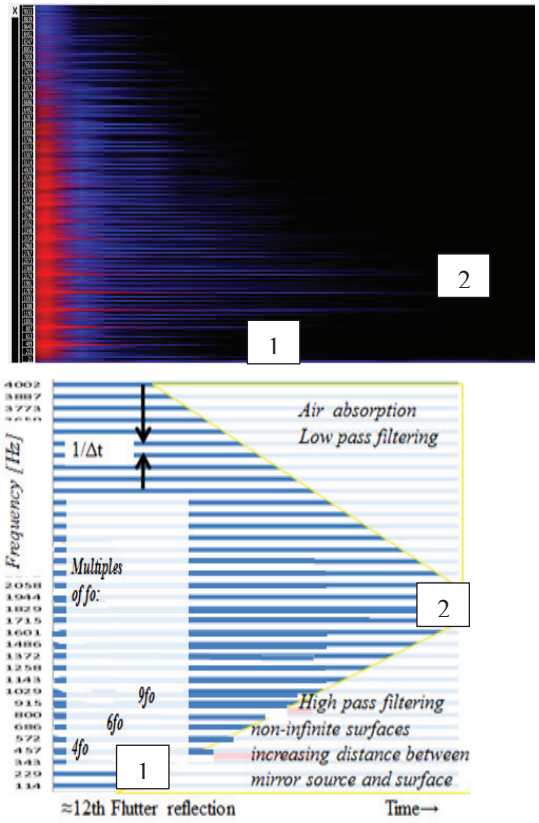


Figure 8. The two main "toneties" of flutter. Measurement and schematic overview

7. FLUTTER AS A SOUND EFFECT

For the composition FLUTR [9], the author used flutter as a major sound effect, of course for rhythmic effects, but also regarding timbre. Because Kuhl´s equations are given for reverberation time, and thus not directly a signal processing algorithm, several typical input parameters were chosen, and used in patches both in Max and Pure Data, and also transferred to plug-ins in Reaper.

doi:10.25836/sasp.2019.37

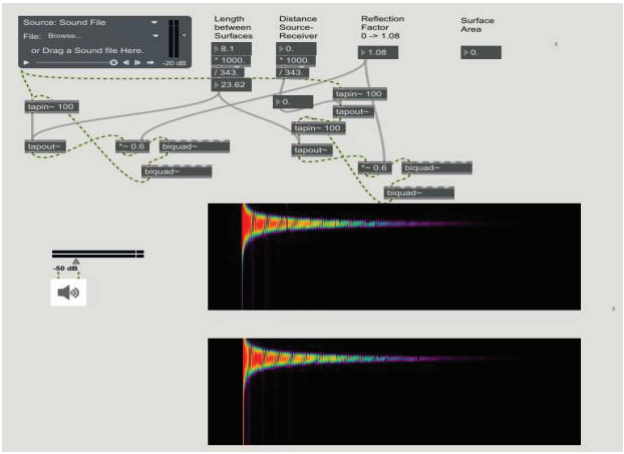


Figure 9. Extract from Max/msp patch for flutter echoes

To check the very simple patch, a Dirac pulse was used as signal and the result in fig. 10 shows good agreement with the measurements shown in the fig.1 and 2.

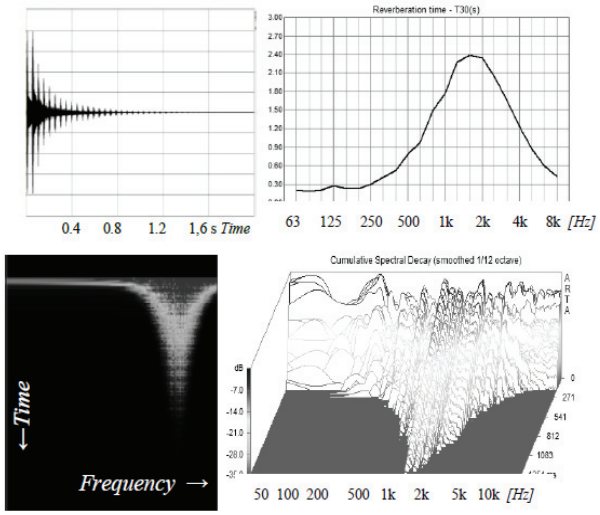


Figure 10. Dirac pulse sent through Max/Msp patch. Impulse Response, Reverberation Time, Spectrogram and Waterfall.

Tests indicate that for "common sized rooms", an exact calculation of the peak frequency of the "flutter tail" is not really necessary. Both calculations using Kuhl´s equations and from the measurements in fig. 6, shows that for a quick simulation, it is sufficient to choose app. 1.5-2 kHz as the center frequency of the band pass filter in the loop. Another reason for not needing to be very precise in the calculation of the frequency of the "tail" is that the frequency is so high that we are over the common melodic range of frequencies.

8. CONCLUSION

This paper shows that the resulting characteristic mid/high frequency timbre of a flutter in ordinary rooms is not a "tone", but a gradual band pass filtering of the broad banded impulsive signal, like a gradual subtractive synthesis. This filtering is a combination of two filtering effects: Low frequency dampening due to the increasing source distance and diffraction, which gives that the sound field is transferred from spherical to plane waves, and High frequency dampening due to air absorption. The result is the characteristic "tonety" mid/high frequency character of a flutter echo, as a "tail", typically ending around 1-2 kHz. Flutter should not be explained by room resonances/ repetition "Tonety", (but might actually be considered as a gradually filtering towards a very high harmonic of this).

We started out stating that flutter echoes had nothing to do with room resonances, but since the fluttering is filtering, the final "target" of the flutter tail actually will be the n^{th} harmonic of the room resonance, (where n might be typically in the order of 50-150). A more detailed simulation for which of these resonances that "win" could be interesting, but for a "real time" simulation, 2 kHz clearly gives the timbre of flutter. For practical use in room simulations and electro-acoustic music, it is found that a very simple way to simulate flutter echoes is to make repetitive repetitions between the two actual dimensions and make a gradual bandpass filtering to about 2 kHz. This has been implemented in Max/Pd and as a plug in.

9. REFERENCES

[1] T. Halmrast, "Why do Flutter always end up around 1-2 kHz?," *Proceedings of the Institute of Acoustics, Room Acoustics, Paris 2015, p. 395-408, paper 53.*

[2] T. Halmrast, "Flutter Echoes; Timbre and possible use as a sound effect," *Proc. of the 18th Int. Conference on Digital Audio Effects (DAFx-15), Trondheim, Norway, 2015*, pp. 213-218.

[3] Odeon Room Acoustics Software. http://www.odeon.dk

[4] D. Y. Maa, "The flutter echoes". *J. Acoust. Soc. Amer.* 13 [1941], 170

[5] E. Krauth, R. Bücklein, "Modelluntersuchungen an Flatter-echos" *Frequenz. Zeitschrift für Schwingungs- und Schwachstromtechnik,* Band 18 Aug. 1964. Nr. 8. pp. 247-252.

[6] W. Kuhl, "Nachhallzeiten schwach gedampfter geschlossener Wellenzüge", *Acustica,* Vol. 55 1984, pp. 187-192

[7] P. Svensson, *"Edge diffraction Matlab toolbox."* http://www.iet.ntnu.no/~svensson/software/ 2013

[8] J.H. Rindel, "Attenuation of sound reflections due to diffraction", *Nordic Acoustical Meeting,* Aalborg, Denmark Aug. 1986, pp. 257-260

[9] T. Halmrast, "The Timbre of Flutter Echoes and the Composition FLUTR", *Int. Computer Music Conference, Proceedings,* June 2019, New York *(video-presentation: www.tor.halmrast.no)*

PERCEPTUAL COMPARISON OF AMBISONICS-BASED REVERBERATION METHODS IN BINAURAL LISTENING

Isaac Engel[1] **Craig Henry**[1] **Sebastià V. Amengual Garí**[2]
Philip Robinson[2] **David Poirier-Quinot**[3] **Lorenzo Picinali**[1]
[1] Dyson School of Design Engineering, Imperial College London, United Kingdom
[2] Facebook Reality Labs, Redmond, US
[3] Sorbonne Université, CNRS, Institut Jean Le Rond d'Alembert, France
isaac.engel@imperial.ac.uk

ABSTRACT

Reverberation plays a fundamental role in the auralisation of enclosed spaces as it contributes to the realism and immersiveness of virtual 3D sound scenes. However, rigorous simulation of interactive room acoustics is computationally expensive, and it is common practice to use simplified models at the cost of accuracy. In the present study, two subjective listening tests were carried out to explore trade-offs between algorithmic complexity (and approach) and perceived spatialisation quality in a binaural spatialisation context. The first experiment assessed the perceived realism of room reverberation, comparing auralisations based on Ambisonic impulse responses at varying resolutions (zeroth to fourth order). The second experiment focused on the perceptual relevance of different approaches to binaural reverb rendering, looking at statically or dynamically rendered room simulations. Throughout these experiments, the direct sound path was rendered separately by convolution with a Head Related Impulse Response (HRIR). Preliminary results suggest that, for the conditions under test, there may not be perceivable benefits in using high order Ambisonics encoding (beyond first order) for room auralisation and that introducing head-tracking may have little impact as well, as long as the direct sound is rendered with enough accuracy. Further work is outlined with regards to continuing this research with a higher number of participants and more varied tested conditions to clarify the extent to which these conclusions can be made.

1. INTRODUCTION

In the years since digital reverb was first suggested by Schroeder and Logan [1] there has been continual development and improvement in the technology. Until recently, research has been driven predominantly by two industries: acoustic architecture and music, in which offline computation is generally sufficient. However, the tech industry is responding to the ever-growing demand for interactive audiovisual experiences. Dynamic room auralisation is an important aspect of mixed reality experiences as it contributes to the quality of spatialisation [2]. With the recent emergence of portable Virtual and Augmented Reality, which have a limited computational power compared to desktop computing, lightweight room auralisation and acoustic modelling are essential. It is therefore important to establish the perceptual relevance of different spatial, temporal and spectral attributes of acoustic room responses to identify where computational savings can be made.

© Isaac Engel, Craig Henry, Sebastià V. Amengual Garí, Philip Robinson, David Poirier-Quinot, Lorenzo Picinali. Licensed under a Creative Commons Attribution 4.0 International License (CC BY 4.0). **Attribution:** Isaac Engel, Craig Henry, Sebastià V. Amengual Garí, Philip Robinson, David Poirier-Quinot, Lorenzo Picinali. "Perceptual comparison of Ambisonics-based reverberation methods in binaural listening", 1st EAA Spatial Audio Signal Processing Symposium, Paris, France, 2019.

2. BACKGROUND AND MOTIVATIONS

Reverberation is the result of the pairing of an acoustic source with an environment. As the wave propagates from the source it interacts with its surroundings. The geometry of the space causes the wave to reflect and diffract. As a result, the wave arrives at the receiver via different paths. Sound has a sufficiently slow speed that these arrivals occur at perceptually distinct time intervals. As the wave continues to interact with the room, higher order reflections decrease in amplitude and the echo density increases, eventually resulting in a diffuse reverberant sound field. A room impulse response (RIR) is typically decomposed into three sections: direct sound, early reflections and late reverberation. The geometry of the room and acoustic properties of the materials therein affect a number of objective acoustic characteristics.

Drawing parallels between objective and subjective spatial features has long been the aim of different avenues of acoustic research. Current literature states that: (i) Direct sound reaches the listener first, allowing the listener to localise the source. (ii) However, strong specular reflections arriving within 5-10ms can contribute to creating a perceived image shift and broaden the apparent source width [3]. (iii) In addition, they colour the timbre of the direct sound due to phase cancellations and subsequent comb-filtering [4]. These reflections are not localised as a separate event due to the precedence effect [5]. (iv) Kaplanis et al. [6] state that the time delay between the direct

doi:10.25836/sasp.2019.11

sound and the first perceptually distinct early reflection affects the perception of presence and the environment dimensions, and that (v) the temporal characteristics of subsequent early reflections also contribute to the environment size. (vi) The timing, direction and spectrum of early lateral reflections contribute to the envelopment of the space [7]. As the density of the reflections increases, perception is governed less by temporal characteristic and more by statistical properties of the reverberant tail. (vii) Research done by Yadav et al. [8] suggests that the reverberation time contributes to the perception of size most significantly in large rooms, whereas early reflections are of greater importance in small rooms. Further observations have been made with respect to binaural rendering. (viii) Reverberation improves the externalisation of sound sources in the binaural domain, even if only early reflections are used [2]. (ix) The congruence between the room acoustic properties of the listening space and those of the presented virtual sounds affect the level of externalisation [9].

When modelling these features for a real time interactive environment, it is the early reflections which present the greatest difficulty, as literature suggests that in order to realistically render a room, each early reflection must be properly spatialised. Otherwise, perceptually relevant characteristics such as envelopment, room dimensions and presence would be inaccurately represented in the resultant signal. On the contrary, the reverberant tail is characterised by statistical properties, individual reflections cannot be differentiated and are uniformly distributed around the listener in their direction of arrival.

Rigorous rendering of head-tracked early reflections generally has a high computational cost. There are two primary solutions: convolution-based reverb and algorithmic reverb. The former relies on convolution with premeasured RIRs which are dynamically swapped with head movements. Whereas, algorithmic reverb attempts to approximate the reverb by various means [10] [11] [12] and, when rendered dynamically, filters must be recomputed in real time. In practice it is common to use simplified models at the cost of accuracy. Therefore, it is relevant to investigate the extent to which computations can be simplified (e.g. by reducing the spatial resolution) without losing perceived spatialisation quality. This paper will concentrate on convolution-based methods, however, conclusions may also be relevant in algorithmic reverb.

3. PREVIOUS WORK

Prior to this paper a study was carried out assessing the realism and localisation accuracy achieved with five different binaural reverberation rendering methods [13]. The primary method was an approach which will be referred to as Reverberant Virtual Loudspeaker decoding (RVL) method.

3.1 RVL: an efficient binaural reverberation method

The RVL method is used to efficiently produce dynamic convolution-based binaural reverberation by using a set of Binaural Room Impulse Responses (BRIRs). The method is inspired by the classic virtual loudspeaker approach, first outlined by McKeag and McGrath [14] and later used by Noisternig et al. [15]. In the original method, one or more sound sources are encoded in the Ambisonic domain, which is then decoded to a set of virtual loudspeakers distributed around the listener. The binaural signal is then obtained by convolving each loudspeaker feed with the HRIR corresponding to its location. RVL follows the same approach, however it uses BRIRs in the loudspeaker position instead of HRIRs, effectively integrating the acoustics of the room in the binaural rendering.

This method has two limitations. The first being that the early reflections for a given source-receiver pair are approximated by the reflections of the limited set of loudspeaker positions. The second is that, when using dynamic rendering, the relative position of the sound sources can be changed in the Ambisonics domain, but the room is head-locked due to the set of BRIRs being fixed. This means that a rotation of the listener's head is equivalent to all of the sources in the room rotating in the opposite direction. This would be irrelevant in an anechoic scenario (traditional virtual loudspeaker method) but not in this case.

The main advantage of the RVL method over other convolution-based approaches is that, as all sound sources are encoded in the same Ambisonic soundfield, the number of required convolutions is independent from the number of sources. This is a big advantage when several sources need to be dynamically rendered at the same time. Furthermore, because the decoding process from Ambisonic channel to the left and right binaural channels is linear and time-invariant, the direct transfer functions between them can be calculated, reducing the total number of convolutions needed in real time to 2 per Ambisonic channel (e.g. 8 convolutions for first order, 18 for second order, etc.). These simplifications create a significant computational saving over other methods (convolution-based or otherwise), as the room auralisation is inherent in the rendering process, instead of having to rely on additional steps to incorporate reverb. Whether the simplifications have a negative impact on the perceived quality of spatialisation, is the question that will be addressed in the present study.

3.2 Previous study: method and results

In the previous study, a subjective listening test was carried out to evaluate the impact of the spatial resolution of reverberation in the perceived quality of binaural rendering [13]. The tested reverberation methods ranged from higher resolution and complexity (RVL with third order Ambisonics and 20 virtual loudspeakers) to less complex solutions (stereo). It also included simplistic approaches such as diotic monophonic reverb and 'mono panned', where the reverb of each sound source was added to the direct sound and spatialised in the same position. For all tested conditions, the direct sound path was rendered by convolving the dry audio with an anechoic HRIR, and was identical across conditions. The BRIRs for the different virtual loudspeaker positions were simulated by means of geometrical acoustic modelling software, without computing the

doi:10.25836/sasp.2019.11

direct sound path.

In the listening test, subjects were presented with all possible pairs of conditions and were asked to compare them in terms of plausibility and spatialisation accuracy. The audio scene included a female voice and footstep sounds moving slowly around the listener. The test was implemented in a online platform in order to reach a large population (75 participants). Results showed that no significant differences were present between four of the five tested methods, with 'mono panned' performing the worst. This seems to suggest that the level of complexity in the reverberation method does not always yield perceptually relevant improvements.

Due to this early test being web-based, this study had some limitations, such as the renderings being static (no head tracking) and the lack of control over the headphones being used by the participants and the playback level. As such, broader conclusions were unable to be made unless these issues were resolved.

4. CURRENT EXPERIMENTS

The literature review in section 2 suggests that early reflections must be properly spatialised to recreate perceptually relevant room characteristics. However the study described in section 3 suggests that increased algorithmic complexity above a certain threshold may no longer lead to improved perceived realism.

As opposed to the aforementioned web-based approach, the current study aimed to carry out experiments under laboratory conditions with dynamic (head-tracked) binaural rendering. Furthermore, reverberation based on recorded IRs is employed instead of geometrical modelling.

Two experiments are outlined in this paper, both of which were designed to test the perceptual limits of binaural rendering complexity. The methods used reproduce early reflections and tails with differing levels of accuracy to test participants' ability to discern between them and state subjective preference on the basis of perceived realism. The questions that these experiments try to answer are:

1. What is the perceptual impact of decreasing the Ambisonic order of reverberation, if direct sound is still rendered accurately? (first experiment)

2. What is the perceptual impact of rendering reverberation statically or dynamically, and using a simplified approach such as RVL? (second experiment)

4.1 Measurements

A small meeting room (RT = 900 ms) was measured for this study in two different ways: (i) RIRs were recorded with an Eigenmike microphone (MH Acoustics [1]) from three directions (-30/0/30° azimuth, 0° elevation) and then encoded into first-fourth order Ambisonics using the EigenStudio software package. (ii) BRIRs were recorded with a KEMAR head and torso simulator from six directions (front, back, left, right, up, down). All measurements were made using the sine sweep technique [16] and a Genelec 8030 loudspeaker at a distance of 1.2 m.



Figure 1: Picture of the room used in this study, before measuring RIRs with the Eigenmike.

The key to this study was that the direct sound path was always rendered separately from the reverberation, which allowed for individual control. To that end, all RIRs and BRIRs had the direct sound path removed by replacing the first 3.71 ms after the onset of all impulse responses with silence and applying a Hanning window – this time was calculated by subtracting the direct sound propagation time from the first reflection (floor) propagation time, minus a safety window of 30 samples (0.68 ms).

Given that the direct sound was obtained from a separate system and database (see next subsection), all impulse responses had to be normalised to match the direct-to-reverberant ratio of the actual room. Furthermore, the Ambisonic RIRs were equalised with a second order low-shelf filter at a gain of -15 dB and a cutoff frequency of 1 kHz to match the perceived coloration of a KEMAR BRIR.

4.2 Audio material and binaural rendering

The stimulus used in the test was an extract of female English speech from the 'Music for Archimedes' corpus [2]. The sound source was spatialised at -30° azimuth, 0° elevation and a distance of 1.2 m.

The direct sound was always rendered through convolution with the Head Related Impulse Response (HRIR) of a KEMAR from the SADIE II database [3]. To generate the reverberation, two different methods were used:

1. **Ambisonic reverb**, convolving the source with the Ambisonic RIRs, and decoding the soundfield into virtual loudspeakers which are then rendered binaurally using HRTFs.

2. **Reverberant Virtual Loudspeaker decoding (RVL) reverb**, using the recorded KEMAR BRIRs.

[1] mhacoustics.com/products

[2] pcfarina.eng.unipr.it/Public/Aurora_CD/Anecoic/Archimedes/CD-cover/Archimedes.htm

[3] york.ac.uk/sadie-project/database.html

doi:10.25836/sasp.2019.11

The 3D Tune-In Toolkit [17] was used as the spatial audio engine. It binaurally rendered both the direct sound and virtual loudspeakers, and allowed for head-tracking using an EdTracker Pro Wireless [4] .

4.3 Experiment 1: MUSHRA test

One of the ways to reduce the number of convolutions required in Ambisonics-based reverb is to decrease the Ambisonic order. However, this has been shown to deteriorate localisation of direct sources [18] [19]. It was the aim of this experiment to assess whether this degraded spatialisation accuracy was also perceptually relevant to the reverberation process, with the direct sound being rendered through a full set of HRIRs.

Listeners were asked to rate the quality of six conditions where reverberation was rendered using differing Ambisonics orders and virtual loudspeaker configurations:

- **Dry** – anechoic direct sound.
- Zeroth order Ambisonics (**0OA**) with 6 virtual loudspeakers in a tetrahedron setup, each playing the W channel.
- First order Ambisonics (**1OA**) with 6 virtual loudspeakers in a tetrahedron setup.
- Second order Ambisonics (**2OA**) with 12 virtual loudspeakers in an icosahedron setup.
- Third order Ambisonics (**3OA**) with 20 virtual loudspeakers in a dodecahedron setup.
- Fourth order Ambisonics (**4OA**) with 32 virtual loudspeakers in a Pentakis-dodecahedron setup.

The six stimuli were presented in a MUSHRA listening test format [20] implemented in MaxMSP, with the **4OA** method acting as the reference, and the **Dry** stimuli as the anchor. Subjects were asked to rate the similarity of each stimulus to the reference on a scale from 0 to 100, where the latter means 'identical to the reference'. In addition, the user interface showed a picture of the simulated room and a diagram with the relative position of the sound source, to assist the subjects in creating an internal reference of the scenario being rendered. Listeners were encouraged to use head movements to explore the scene.

4.4 Experiment 2: Paired comparisons

The second experiment focused on the perceptual relevance of different approaches to binaural reverb rendering. Three different rendering methods were compared:

- Ambisonic reverb, with first order Ambisonics and 6 virtual loudspeakers – i.e. the **1OA** method from the MUSHRA test.
- **1OAS**: Static (non-head-tracked) Ambisonic reverb, with first order Ambisonics and 6 virtual loudspeakers – direct sound path was still tracked.
- First order **RVL** reverb obtained from the KEMAR BRIRs, which is inherently "static" in the sense that the room is head-locked as explained in section 3.1.

The two key factors explored in the second experiment are (i) the importance of the sound field being rendered statically or dynamically and (ii) the perceptual relevance of accurate early reflections. The three tested methods approximate early reflections to different extents by making various simplifications. In theory, this should interfere with most of the perceptual attributes outlined in section 2 as it alters the spatial, temporal and spectral characteristics of the room. Whether the differences between these approaches are perceptually relevant is still to be understood.

In each trial, listeners were shown a picture of the rendered room and a diagram with the position of the sound source, and were presented two stimuli, A and B. The question asked was '*Considering the given scene, which example is more appropriate?*'. The rating scale was continuous from -2 to +2, with one decimal place, from *Definitely A* to *Definitely B* (see Fig. 3). Listeners were allowed to freely switch between the synchronised stimuli during a trial, and head movements were encouraged to explore the scene. All possible pairs of the three rendering methods were tested, plus two null pairs where A and B were identical (randomly chosen), totalling 8 trials for each subject.

5. INITIAL RESULTS

At the time of writing this paper, the study is still ongoing and only preliminary data are reported. Results for the first five subjects (ages 23-40, 1 female) are presented.

5.1 MUSHRA test

The results of the MUSHRA test are shown in Fig. 2. Descriptive analysis showed that the mean rating for **Dry** results is clearly the lowest, followed by **0OA**, while the other four methods had relatively similar mean ratings.

Due to the low amount of subjects available so far, non-parametric statistical analysis was used. Friedman test showed that the differences between methods is significant ($\chi^2(5) = 20.28, p = 0.001$). At this early stage, post-hoc test results are not reported as are not likely to be reliable due to the low participant count. However, inspecting the boxplot diagram it seems clear that the significant differences are likely between **Dry** and the rest, and between **0OA** and the rest.

5.2 Paired comparisons

Figure 3 shows the comparison ratings for every pair of stimuli. Descriptive analysis showed that mean rating was close to zero for the null pairs, that listeners tended to favour **1OA** and **1OAS** over **RVL**, and that **1OA** and **1OAS** were perceived as very similar, with a slight trend towards favouring the latter.

As done for the other test, non-parametric statistical analysis was used; the Friedman test showed that the differences between the tested pairs were not significant

[4] edtrackerpro.mybigcommerce.com/edtracker-pro-wireless/

doi:10.25836/sasp.2019.11

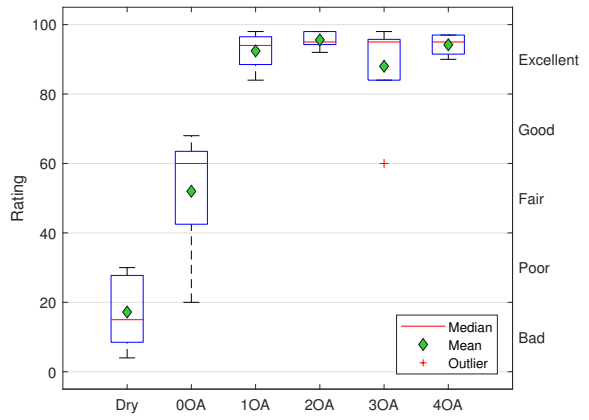


Figure 2: Boxplot diagram of the ratings for each of the tested conditions in the MUSHRA test.

($\chi^2(3) = 6.21, p = 0.102$). Furthermore, one-sample t-tests showed that none of the pairs were significantly different from a normal distribution with mean equal to zero.

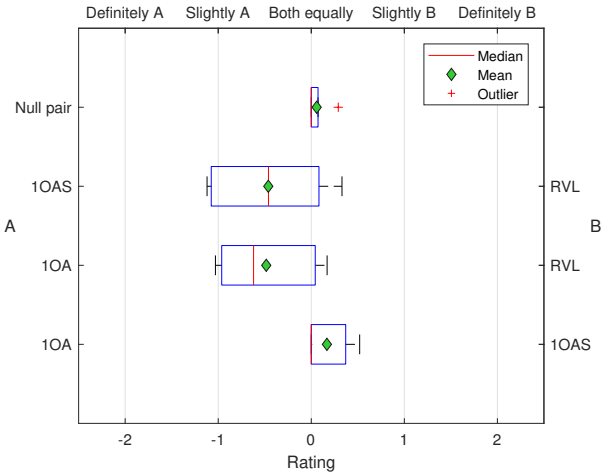


Figure 3: Boxplot diagram of the paired comparison ratings. Negative ratings indicate preference for the stimulus on the left (A), and positive ratings for the stimulus on the right (B).

6. DISCUSSION AND FURTHER WORK

This paper presents early stage results with limited scope considering the range of variables which could affect the outcome of this experiment, namely the program content (i.e. spatialised signal), room size and direct to reverberant ratio, and the participant's level of listening expertise.

With this in mind, the results from the MUSHRA test showed that while listeners gave consistently low ratings to the **Dry** and **0OA** conditions, they were not able to reliably distinguish between first order Ambisonics and above. This result suggests that the findings of [18] [19] regarding improvements in spatialisation in higher orders of Ambisonics may not extend to the reverberant portion of room auralisation, for the conditions tested so far in this study.

The implication of this would be that, provided the direct sound is rendered with sufficient accuracy, the perceptual impact of reverberant spatial resolution may not be as high as expected. Further work (currently being carried out) includes collecting data from additional participants (including a number of expert listeners) and using a range of stimulus types and room shapes and sizes.

A relevant matter that may have influenced these results is that, due to the listening test taking place outside of the measured room, room divergence is masking differences in the performance of the different Ambisonic orders [9]. However, due to the MUSHRA test focusing on perceived differences against a reference, this seems unlikely. In order to eliminate this potential variable, it will be considered to conduct future tests inside the rooms corresponding to the measured impulse.

In the second experiment, no statistically significant differences were found, but the trend suggests that listeners slightly favoured Ambisonic reverb over the RVL method. This may indicate that the simplifications made by RVL could have had a negative, yet not significant, impact in the perceived quality of the reverb. However, it could also be that timbral differences due to the methods being generated from different microphones' recordings – only partially mitigated by heuristic equalisation – were culpable instead. Additionally, listeners were exposed to the MUSHRA test where the reference was an Ambisonic reverb, and may have been biased towards this technique and its timbral quality.

Interestingly, **1OAS** was rated similarly to **1OA**. This is a surprising result, as it suggests that rendering reverb statically instead of dynamically may not be perceptually relevant, assuming that the direct path is rendered through convolution with an HRIR. This goes against considerable literature supporting the contrary, and the scope of these findings is currently limited. The number of participants is still low, and only one room and one stimulus type have been tested. More participants and more varied test conditions are needed to make such claims. Additionally, whilst listeners were encouraged to move their heads, in future work this should be considered with greater detail, either by tracking the movement of participants or prescribing specific head rotations to be carried out.

7. CONCLUSIONS

In this study, the issue of the trade-off between computational complexity and perceived quality for binaural reverberation has been addressed. Preliminary results of a perceptual listening test suggest that the Ambisonic order and the use of head tracking have little perceivable effect on Ambisonic-based reverb, assuming that the direct sound path rendering is accurate enough. Further work has been outlined in the Discussion section.

8. ACKNOWLEDGEMENTS

This work was partly supported by the PLUGGY project (https://www.pluggy-project.eu/), European Unions Hori-

zon 2020 research and innovation programme under grant agreement No 726765.

9. REFERENCES

[1] M. R. Schroeder and B. F. Logan, "'colorless' artificial reverberation," *J. Audio Eng. Soc*, vol. 9, no. 3, pp. 192–197, 1961.

[2] D. R. Begault, E. M. Wenzel, and M. R. Anderson, "Direct Comparison of the Impact of Head Tracking, Reverberation, and Individualized Head-Related Transfer Functions on the Spatial Perception of a Virtual Speech Source," *Journal of the Audio Engineering Society*, vol. 49, pp. 904–916, oct 2001.

[3] S. E. Olive and F. E. Toole, "The detection of reflections in typical rooms," *Journal of the Audio Engineering Society*, vol. 37, no. 7, pp. 539–553, 1989.

[4] S. Bech, "Timbral aspects of reproduced sound in small rooms. II," *The Journal of the Acoustical Society of America*, vol. 99, no. 6, pp. 3539–3549, 1996.

[5] H. Wallach, E. B. Newman, and M. R. Rosenzweig, "The precedence effect in sound localization," *The American Journal of Psychology*, vol. 62, no. 3, pp. 315–336, 1949.

[6] N. Kaplanis, S. Bech, S. H. Jensen, and T. van Waterschoot, "Perception of reverberation in small rooms: a literature study," in *Audio Engineering Society Conference: 55th International Conference: Spatial Audio*, pp. 1–14, Audio Engineering Society, 2014.

[7] M. Barron and A. H. Marshall, "Spatial impression due to early lateral reflections in concert halls: the derivation of a physical measure," *Journal of Sound and Vibration*, vol. 77, no. 2, pp. 211–232, 1981.

[8] M. Yadav, D. A. Cabrera, L. Miranda, W. L. Martens, D. Lee, and R. Collins, "Investigating auditory room size perception with autophonic stimuli," in *Audio Engineering Society Convention 135*, p. 10, Audio Engineering Society, 2013.

[9] S. Werner, F. Klein, T. Mayenfels, and K. Brandenburg, "A summary on acoustic room divergence and its effect on externalization of auditory events," in *2016 Eighth International Conference on Quality of Multimedia Experience (QoMEX)*, pp. 1–6, IEEE, June 2016.

[10] J. C. Allred and A. Newhouse, "Applications of the monte carlo method to architectural acoustics," vol. 30, no. 1, pp. 1–3.

[11] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating smallroom acoustics," vol. 65, no. 4, pp. 943–950.

[12] E. De Sena, H. Haciihabiboglu, Z. Cvetkovic, and J. O. Smith, "Efficient synthesis of room acoustics via scattering delay networks," vol. 23, no. 9, pp. 1478–1492.

[13] L. Picinali, A. Wallin, Y. Levtov, and D. Poirier-Quinot, "Comparative perceptual evaluation between different methods for implementing reverberation in a binaural context," in *Audio Engineering Society Convention 142*, pp. 1–13, 2017.

[14] A. Mckeag and D. Mcgrath, "Sound Field Format to Binaural Decoder with Head Tracking," *AES 6th Australian Regional Convention*, 1996.

[15] M. Noisternig, T. Musil, A. Sontacchi, and R. Holdrich, "3d binaural sound reproduction using a virtual ambisonic approach," in *IEEE International Symposium on Virtual Environments, Human-Computer Interfaces and Measurement Systems*, pp. 174–178, IEEE, 2003.

[16] A. Farina, "Advancements in impulse response measurements by sine sweeps," in *Audio Engineering Society Convention 122*, Audio Engineering Society, 2007.

[17] M. Cuevas-Rodríguez, L. Picinali, D. González-Toledo, C. Garre, E. de la Rubia-Cuestas, L. Molina-Tanco, and A. Reyes-Lecuona, "3d tune-in toolkit: An open-source library for real-time binaural spatialisation," *PLOS One*, vol. 14, no. 3, 2019.

[18] E. Bates, G. Kearney, D. Furlong, and F. Boland, "Localization accuracy of advanced spatialisation techniques in small concert halls," *The Journal of the Acoustical Society of America*, vol. 121, no. 5, pp. 3069–3070, 2013.

[19] A. Sontacchi, P. Majdak, M. Noisternig, and R. Höldrich, "Subjective Validation of Perception Properties in Binaural Sound Reproduction Systems," *AES 21st International Conferece*, pp. 1–4, 2002.

[20] "Itu-r bs. 1534-3: Method for the subjective assessment of intermediate quality level of audio systems," October 2015.

 doi:10.25836/sasp.2019.11

A LINEAR PHASE IIR FILTERBANK FOR THE RADIAL FILTERS OF AMBISONIC RECORDINGS

C. Langrenne[1], É. Bavu[1] and A. Garcia[1]
[1] Conservatoire National des Arts et Métiers (CNAM)
Laboratoire de Mécanique des Structures et des Systèmes Couplés (LMSSC)
christophe.langrenne@lecnam.net

ABSTRACT

Higher order ambisonics involve excessive bass boosts, especially for high orders and at low frequencies. In order to avoid unnecessary and excessive amplification for components that do not contribute significantly to the soundfield, a filterbank can be used in order to cut-off noise amplification. In the present paper, we propose a linear phase IIR filterbank implementation that allows to avoid the used of fast block convolutions or nonlinear IIR filters. Our approach is based on local overlap-add time reversed block convolutions, which allow the filterbank to exhibit a linear delay, which only depends on the sectioned block size. When combined with radial filters of a rigid spherical recording array, this approach allows to change the cutoff frequencies of the filterbank with more flexibility than precomputed FIR filterbanks.

1. INTRODUCTION

Higher order Ambisonics decomposition of natural soundfields is often performed using rigid spherical microphone arrays, mainly because of its simple implementation [1, 2]. All the electronic equipment can be conveniently placed inside the spherical measurement array, without affecting the scattered acoustic field. However, restitution systems for HOA sound field synthesis generally exhibit a much larger radius than measurement arrays. The well-known bass-boost effect is directly linked to the relatively small size of the measurement array: low frequencies have to be amplified, especially for higher order components of the Ambisonics decomposition. The dynamic range for filtering purposes is limited, mainly by the signal-to-noise ratio of the microphone array. In order to overcome this problem, we developed a microphone array prototypes using analog MEMS microphones, which have become a viable solution in a small packaging, with a reasonable price, thanks to the growing use of these sensors in domotics and in the mobile phone industry. MEMS microphone from the same production batch exhibit very similar characteristics and can be used for array signal processing without any level or phase calibration. The proposed prototype is made of groups of 4 MEMS microphones for the same sensor position, in order to improve the signal to noise ratio by 6 dB. This microphone array is a 5^{th} order Ambisonics system (50 sensors 200 MEMS on a Lebedev grid).

Nevertheless, this approach does not dispense from the need to filter higher order coefficients. A simple high-pass filtering on each order component is not sufficient, since this would not only cause losses in terms of amplitude and power but would also affect the loudness of the restitution. A filterbank is therefore needed in order to cut-off noise amplification at low frequencies, and to apply appropriate gains for loudness equalization. For this purpose, Baumgartner *et al.* [4] proposed a non-linear phase filterbank based on Linkwitz-Riley IIR filters. In order to avoid group delay distortions, Zotter proposed a linear phase filterbank based on FIR filters and the use of fast block convolutions [5]. This solution is although not very flexible, since the FIR strongly depend on the radius of the measurement array, and on the filterbank's cut-off frequencies. Any change in the measurement system require a new computation of each corresponding FIR filters.

In the present paper, a linear phase IIR filterbank is implemented. Thanks to the use of local overlap-add time reversal blocks [6], the filterbank exhibits a linear phase delay which only depends on the the time reversal blocksize. The proposed filterbank implementation allows to change in real-time the frequency bands and the loudness equalization (diffuse or free field equalization) using the Faust programming language [7].

2. AMBISONICS

In this paper, we use the following convention for spherical coordinates:

$$\begin{cases} x = r\cos(\theta)\cos(\delta) \\ y = r\sin(\theta)\cos(\delta) \\ z = r\sin(\delta) \end{cases} \quad (1)$$

with r, θ and δ denoting the radius, the azimuth and the elevation angle (see Fig. 1).

© C. Langrenne[1], É. Bavu[1] and A. Garcia[1]. Licensed under a Creative Commons Attribution 4.0 International License (CC BY 4.0). **Attribution:** C. Langrenne[1], É. Bavu[1] and A. Garcia[1]. "A linear phase IIR filterbank for the radial filters of ambisonic recordings", 1st EAA Spatial Audio Signal Processing Symposium, Paris, France, 2019.

doi:10.25836/sasp.2019.03

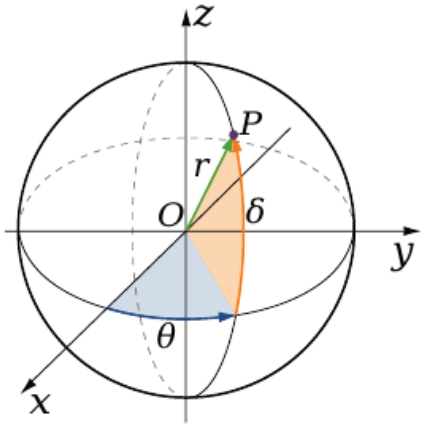


Figure 1. Spherical coordinate system. A point P (x, y, z) is described by the radius r, azimuth θ and elevation δ.

For any incident sound field, the pressure $p(kr, \theta, \delta)$ can be approximated as a truncated Fourier-Bessel series of order M:

$$p(kr, \theta, \delta) = \sum_{m=0}^{M} i^m j_m(kr) \sum_{n=-m}^{m} B_{mn} Y_{mn}(\theta, \delta) \quad (2)$$

with $i = \sqrt{-1}$, k the wave number, $j_m(kr)$ the spherical Bessel function, $Y_{mn}(\theta, \delta)$ the real spherical harmonic of order m and degree n and B_{mn} the wave spectrum.

In (2), the spherical harmonics are defined by:

$$Y_{mn}(\theta, \delta) = \sqrt{(2m+1)\epsilon_n \frac{(m-|n|)}{(m+|n|)}} P_m^{|n|}(\sin(\delta)) \times \begin{cases} \cos(|n|\theta) & \text{if } n \geq 0 \\ \sin(|n|\theta) & \text{if } n < 0 \end{cases} \quad (3)$$

with $\epsilon_n = 1$ for $n = 0$ and $\epsilon_n = 2$ for $|n| > 0$, $P_m^{|n|}$ are the associated Legendre polynomials of order m and degree $|n|$.

2.1 Encoding and decoding

In this section, we consider both spherical microphone and loudspeaker arrays, whose transducers are sampled on a spatial grid with L nodes calculated from a quadrature rule which is exact up to the order M. The recording, encoding and decoding operations are summarized as follows, in the frequency domain [9]:

$$s_{spk} = \underbrace{W Y F_{spk}^{-1}}_{decoding} \quad \underbrace{E_{mic} Y^T W s_{mic}}_{encoding} \quad (4)$$

where $s_{spk}^{L\times 1} = [s_1, s_2, \cdots, s_l, \cdots s_L]^T$ is the vector of input signals for the loudspeakers, $W^{L\times L}$ is the diagonal matrix of quadrature weights, $Y^{L\times(M+1)^2}$ is the matrix of spherical harmonics evaluated at the different nodes. $F_{spk}^{-1^{(M+1)^2\times(M+1)^2}}$ is the diagonal matrix of near field compensation filters with diagonal term:

$$F_{spk}(kr_{spk}) = i^{-(m+1)} k h_m^{(2)}(kr_{spk}) \quad (5)$$

where r_{spk} is the distance of the loudspeakers and $h_m^{(2)}$ the spherical Hankel function of second order. $E_{mic}^{(M+1)^2\times(M+1)^2}$ is the diagonal matrix of equalization filters for the rigid spherical microphone array with diagonal term given by:

$$E_{mic}(kr_{mic}) = i^{-(m-1)} (kr_{mic})^2 h_m^{'(2)}(kr_{mic}) \quad (6)$$

where r_{mic} is the radius of the rigid sphere supporting the microphones and $h_m^{'(2)}(kr)$ the derivative of the spherical Hankel function of second order with respect to the variable kr.

Finally $s_{mic}^{L\times 1}$ are the signals captured by the microphones. Transposition is denoted by the superscript T.

2.2 Radial Filters

The present paper specifically focuses on the radial filters, which correspond to the correction of the rigid spherical measurement array scattering and to the compensation of the restitution loudspeakers, which is obtained usingEqs. (5) and (6):

$$G_m(kr_{mic}, kr_{spk}) = -kr_{mic}^2 \frac{h_m^{'(2)}(kr_{mic})}{h_m^{(2)}(kr_{spk})} \quad (7)$$

For a digital implementation of these filters, one can use the following expressions [10]:

$$h_m^{(2)}(kr) = i^{m+1} \frac{e^{-ikr}}{kr} \sum_{k=0}^{m} a_{m,k} (ikr)^{(-k)} \quad (8)$$

with

$$a_{m,k} = \frac{(m+k)!}{(m-k)!k!2^k} \quad (9)$$

and

$$h_m^{'(2)}(kr) = \frac{m}{kr} h_m^{(2)}(kr) - h_{m+1}^{(2)}(kr) \quad (10)$$

Using eqs. (7), (8) and (10), the radial filters can be expressed as:

$$G_m(kr_{mic}, kr_{spk}) = ikr_{mic} r_{spk} e^{-ik(r_{mic}-r_{spk})} \frac{\sum_{k=0}^{m+1} a_{m,k} (ikr_{mic})^{(-k)} - m \sum_{k=0}^{m} a_{m,k} (ikr_{mic})^{(-k-1)}}{\sum_{k=0}^{m} a_{m,k} (ikr_{spk})^{(-k)}} \quad (11)$$

In this equation, the term $e^{-ik(r_{mic}-r_{spk})}$ is a delay corresponding to the propagation time between the radius of the loudspeakers and the radius of the rigid sphere and can be omitted. The term ik can be interpreted as a differentiator filter. The last term can be efficiently implemented as a discrete-time IIR filter, as proposed by Daniel for nearfield compensation [11]. In our practical implementation, we use the formation proposed by Adriaensen in [12].

doi:10.25836/sasp.2019.03

3. LINEAR PHASE IIR FILTERBANK DESIGN

At low frequencies and for high order components, the compensation gains exhibit very high values. As a consequence, there is a need to stabilize the filters for different Ambisonic orders m, using high-pass filters, whose slopes exceed $6m$ dB/oct. However, the use of this procedure induces a noticeable loudness loss below each cut-off frequency, due to the filtering of signals of higher orders than m. One way of circumventing this problem consists in designing a filterbank with cross-overs with compensated amplitudes as proposed by Baumgartner in [4], who proposed the use of a Butterworth filterbank with all-pass based phase compensation in order to have the same non-linear phase for each of the decomposition orders.

In order to avoid group delay distortions at low frequencies, we propose a realtime implementation of a linear phase IIR filterbank, whose implementation is based on the Powell and Chau technique [6]. This procedure is based on a conventional two-pass IIR filter. The time reversed section implementation of the noncausal function $H(z^{-1})$ is cascaded with the original causal IIR filter $H(z)$. The equivalent transfert function therefore becomes:

$$H_{eq}(z) = H(z^{-1})H(z)e^{-j\omega d(L)} \tag{12}$$

whose phase is linear and magnitude is equal to the square of the magnitude of the original IIR filter $H(z)$. The term $e^{-j\omega d(L)}$ corresponds to the overall delay induced by the time reversal sectioning procedure, which is $(4L-1)$ samples in our implementation.

Using a Butterworth filter $H(z)$, one obtains a linear Linkwitz-Riley filter, with crossover frequencies located at -6 dB attenuation. Using the well-known overlap-add method for sectioned convolution, the output can be obtained in realtime from a superposition of finite length responses from adjacent input sections of length L. The finite length of these input section correspond to a truncature of the IIR filter, which can cause a ripple in the pass-band section of the filter. Taking a sufficiently long value of L allows to ensure that the response magnitude is undistinguishable from the ideal magnitude response, with a nearly constant group delay.

4. REAL-TIME IMPLEMENTION WITH FAUST

Faust (Functional Audio Stream) is a functional programming language for sound synthesis and audio processing with a strong focus on the design of synthesizers, musical instruments, audio effects, etc. Faust targets high-performance signal processing applications and audio plug-ins for a variety of platforms and standards. It is used on stage for concerts and artistic productions, in education and research, in open source projects as well as in commercial applications.

4.1 Linear phase filter

Table (1) recalls the conventional overlap-add method for sectioned convolution. It is worth noticing that the original signal is split in two branchs in order to separate successive sections of L samples. Each branch is filtered using an IIR filter which is reset at the end of $2L$ samples, because it is a truncated sectioned convolution.

This overlap-add scheme can also be used to implement noncausal time reversed convolutions. The process proposed by Powell and Chau [6] consists in the following (see Table (2)):

1. Time reverse each input section using a LIFO (Last Input First Output)
2. Split the reversed sectioned signal and use the reseted IIR filter on each branch
3. Create output sections consisting of the trailing response from the current input section plus the leading response from the previous input section. To be implemented in real-time, this supposes to split again the branch to separate leading and trailing sections, and to delay the leading section of each branch by $2L$ samples
4. Time reverse the output sections using a LIFO.

4.2 Faust implementation

We use the following Faust implementation of a LIFO:

```
LIFO(L) = @(phase(L)*2) with {
phase(n) = (1) : (+ : %(n)) ~ _ ;
};
```

Each LIFO therefore induces a delay of $(L-1)$ samples before performing the time reversed sections. As a consequence, the clocks used to obtain the different branches begin with a first section of $(L-1)$ samples, and the following clock sections are of length L.

The overall FAUST implementation of a linear phase filter is shown on Fig. 2. A one-sample delay before the second LIFO is used in order to compensate the missing sample of the first LIFO. This therefore allows to use the same clocks, both for forward and backward filterings. Since each LIFO induces a delay of $(L-1)$ samples and the reversed time section induces a delay of $2L$ samples, this linear phase IIR filtering process exhibits an overall delay of $(4L-1)$ samples.

5. LINEAR PHASE FILTERBANK

The proposed filterbank implementation only makes use of passband filters. We made this choice in order to avoid soliciting the loudspeakers at very low frequencies. At high frequencies, a low-pass filter is added in order to combine it with the differentiator filter given in (11). The zero and the pole of these two filters compensate in order to avoid problems at high frequencies, when only the differentiator filter is used. This filter is not present in the first subsection, but only implemented in the second subsection.

doi:10.25836/sasp.2019.03

	section k (L)	section k+1 (L)	section k+2 (L)
x(n) clk_1 clk_2	$x_k(1),\cdots,x_k(L)$ $1\cdots1$ $0\cdots0$	$x_{k+1}(1),\cdots,x_{k+1}(L)$ $0\cdots0$ $1\cdots1$	$x_{k+2}(1),\cdots,x_{k+2}(L)$ $1\cdots1$ $0\cdots0$
x(n) $\times clk_1$ x(n) $\times clk_2$	$x_k(1)\cdots,x_k(L)$ $0\cdots0$	$0\cdots0$ $x_{k+1}(1)\cdots,x_{k+1}(L)$	$x_{k+2}(1),\cdots,x_{k+2}(L)$ $0\cdots0$
IIR(x(n) $\times clk_1$) IIR(x(n) $\times clk_2$)	Leading k Trailing $k-1$	Trailing k Leading $k+1$	Leading $k+2$ Trailing $k+1$
Sum=y(n)	$y_k(n)=L_k+T_{k-1}$	$y_{k+1}(n)=L_{k+1}+T_k$	$y_{k+2}(n)=L_{k+2}+T_{k+1}$

Table 1. Overlap-add method for sectioned convolution

	section k (L)	section k+1 (L)	section k+2 (L)
x(n) clk_1 clk_2	$x_k(1),\cdots,x_k(L)$ $1\cdots1$ $0\cdots0$	$x_{k+1}(1),\cdots,x_{k+1}(L)$ $0\cdots0$ $1\cdots1$	$x_{k+2}(1),\cdots,x_{k+2}(L)$ $1\cdots1$ $0\cdots0$
LIFO(L)=x(-n)	$x_k(L),\cdots,x_k(1)$	$x_{k+1}(L),\cdots,x_{k+1}(1)$	$x_{k+2}(L),\cdots,x_{k+2}(1)$
x(-n) $\times clk_1$ x(-n) $\times clk_2$	$x_k(L)\cdots,x_k(1)$ $0\cdots0$	$0\cdots0$ $x_{k+1}(L)\cdots,x_{k+1}(1)$	$x_{k+2}(L),\cdots,x_{k+2}(1)$ $0\cdots0$
IIR(x(-n) $\times clk_1$) IIR(x(-n) $\times clk_2$)	Leading k Trailing k-1	Trailing k Leading k+1	Leading k+2 Trailing k+1
(IIR(x(-n) $\times clk_1$) $\times clk_1$)@(2L) IIR(x(-n) $\times clk_1$) $\times clk_2$ (IIR(x(-n) $\times clk_2$) $\times clk_2$)@(2L) IIR(x(-n) $\times clk_2$) $\times clk_1$	Leading k-2 $0\cdots0$ $0\cdots0$ Trailing k-1	$0\cdots0$ Trailing k Leading k-1 $0\cdots0$	Leading k $0\cdots0$ $0\cdots0$ Trailing k+1
Sum=y(-n)	$y_{k-2}(-n)=L_{k-2}+T_{k-1}$	$y_{k-1}(-n)=L_{k-1}+T_k$	$y_k(-n)=L_k+T_{k+1}$
LIFO(L)=y(n)	$y_{k-2}(n)$	$y_{k-1}(n)$	$y_k(n)$

Table 2. Overlap-add method for time reversed convolution. @(2L) means to apply a delay of 2L samples.

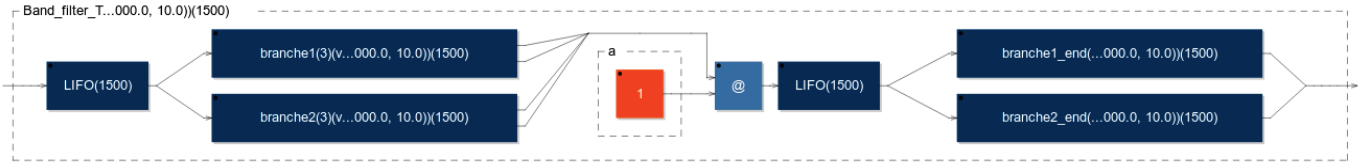


Figure 2. Block diagram of the Faust implemention.

5.1 Without radial filters of ambisonic recording

Fig. 3 shows the filterbank response for a sectioned convolution of length $L = 100$ samples. This figure shows that the choice of $L = 100$ is not sufficient for the lower passband filter, which has a lower bound of 20 Hz. At low frequencies, the impulse response is long and the truncated IIR length must be larger. On the other hand, one can notice that for high passband filters, the dynamic range are > 100 dB with only 100 sample sections.

Fig. 4 shows that the use of larger L values allow to get closer to the ideal filter response and very low frequencies. With $L = 900$ samples, the filterbank response is shown on Fig. 5. Its total response is flat by design, since it corresponds to a Linkwitz-Riley filterbank with crossover frequencies located at -6 dB attenuation.

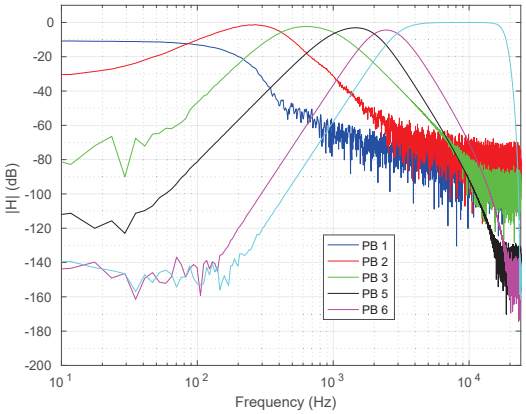


Figure 3. Filterbank response with $L = 100$

doi:10.25836/sasp.2019.03

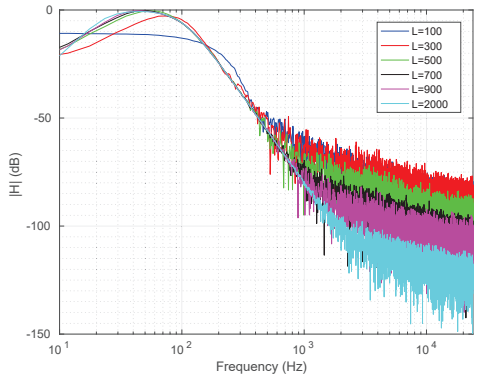


Figure 4. Low passband filter with different lengths of sectioned convolutions.

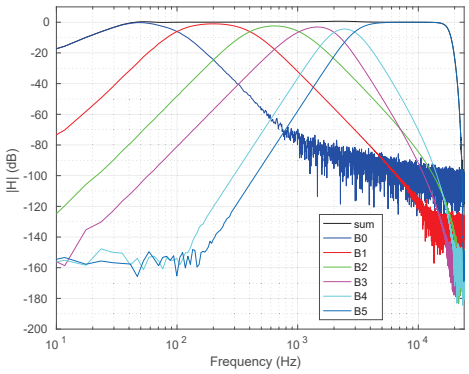


Figure 5. Filterbank response with $L = 900$.

As shown on Fig. 7, the group delay is almost constant around the ideal value of $(4L-1)/F_s$, with $F_s = 48$ kHz. Fig. 7 allows to show that the group delay only deviates from this ideal value by less that 0.5 msec from 20 Hz to 20 000 Hz.

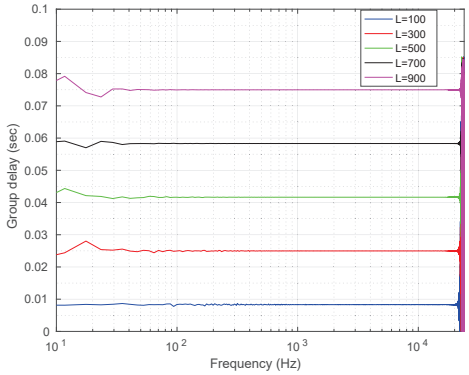


Figure 6. Group delays for different lengths of sectioned convolutions.

5.2 With radial filters of ambisonic recording

Our 5$^{\text{th}}$ order ambisonic recording prototype is a rigid sphere of radius 0.07 m. The 50 MEMS microphones are spatially sampled on the sphere using a Lebedev rule (see Fig. 8). Because of the large radius, this is not necessary to filter the first ambisonic order. In return, the aliasing frequency is 6500 Hz. Our spherical restitution prototype is composed of 50 loudspeakers, on the same Lebedev grid, at a radius of 1.07 m. For the low frequencies, six larger

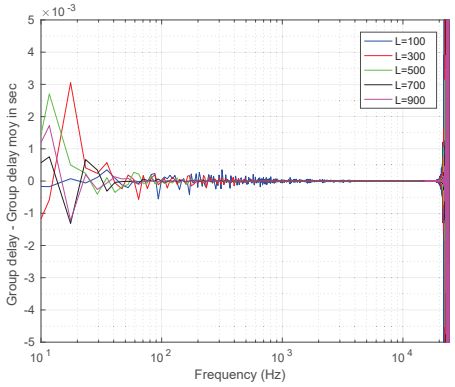


Figure 7. Deviations from the ideal group delays, for different lengths of sectioned convolutions.

loudspeakers are added and driven with ambisonic signals up to order 1 ((see Fig. 9).



Figure 8. Cnam prototype of 5th order ambisonic recording sphere.

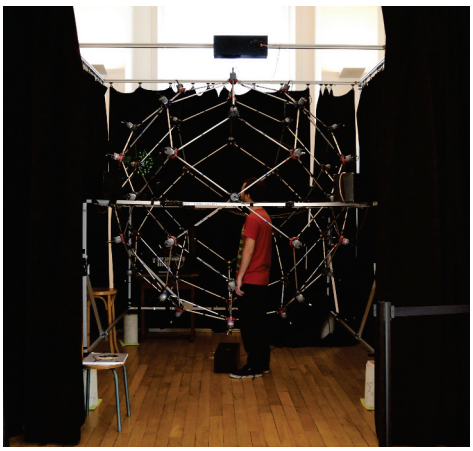


Figure 9. Cnam prototype of 5th order ambisonic restitution sphere.

Fig. 10 and Fig. 11 show the magnitude and the phase responses of the radial filter obtained after the proposed filterbank. The delay term of (11) is not included, and the ideal delay of (4L-1) samples induced by the proposed filterbank implementation is also compensated. When compared with the theoretical filters of (11), small differences

can be noticed above 10 kHz, due to the differentiator filter, which is above the aliasing frequency. The filterbank cutoff frequencies have been chosen in order to prevent any gain values exceeding 40 dB. These values can be changed in real time with a slider in the Faust panel. An amplitude compensation gain has also been added in the final version, as suggested by Baumgartner in [4].

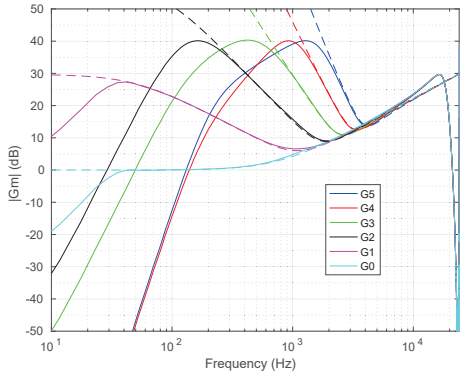


Figure 10. Magnitude response of G_m. Solid line : measured after the filterbank , Dashed line: ideal eq.(11).

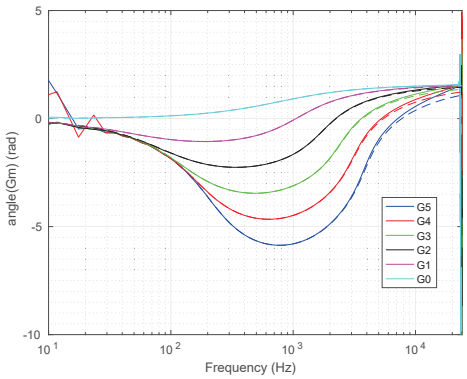


Figure 11. Phase response of G_m. Solid line : measured after the filterbank , Dashed line: ideal eq.(11).

6. CONCLUSIONS

A linear phase IIR filterbank is presented in this paper. When combined with the radial filters of an ambisonic recording rigid sphere, it allows to change the cutoff frequencies in real-time in order to adjust them more precisely, while achieving a linear phase filterbank. In order to meet the requirements for the lower passband filter, a truncated sectioned IIR filter with 900 samples is needed for $F_s = 48$ kHz, , which induces an overall delay of 75 msec.

The final code can be compiled up to the 5th ambisonic order. However, some buffer overflows appear from time to time at the 5th order. The code runs on only one CPU core, and the solution could come from the use of multiple cores, which is a challenge for audio stream synchronization.

7. ACKNOWLEDGMENTS

The authors would like to gratefuly thank Yann Orlarey and Stéphane Letz from the Grame institut. Yann Orlarey helped us with the LIFO Faust code and Stéphane Letz took his time to help us to improve the Faust code compilation. We also express our gratitude to Philippe Chenevez, head of Cinela society, for the realization of the so-called "MemsBedev" ambisonic recording sphere.

8. REFERENCES

[1] S. Bertet, J. Daniel, E. Parizet, L. Gros, and O. Warusfel. Investigation of the perceived spatial resolution of higher order ambisonics sound fields: a subjective evaluation involving virtual and real 3D microphones. In *AES 30th International Conference*, Saariselk, Finland, 2007.

[2] J. Meyer and G. Elko. A highly scalable spherical microphone array based on an orthonormal decomposition of the soundfield, *Proceedings of ICASSP02*, vol 2, Orlando, FL, USA, 2002.

[3] S. Favrot, M. Marschall , J. Ksbach , J. Buchholz, T. Weller. Mixed-order Ambisonics recording and playback for improving horizontal directionality, presented at the *AES 131st convention*, New York, USA, 2011.

[4] R. Baumgartner, H. Pomberger, and M. Frank. Practical Implementation of Radial Filters for Ambisonic Recordings, in *Proc. first International Conference on Spatial Audio*, Detmold, Germany, 2011.

[5] F. Zotter. A Linear-Phase Filter-Bank Approach to Process Rigid Spherical Microphone Array Recordings, in *Proceedings the 5th International Conference on Electrical, Electronic and Computing Engineering, IcETRAN 2018*, Pali, Serbia, 2018.

[6] S.R Powell and P.M Chau. A technique for realizing linear phase IIR filter, in *IEEE transactions and signal processing*, vol 29(11), pages 2425–2435, 1991.

[7] for Faust programming language, see https://faust.grame.fr/.

[8] P. Lecomte and P.-A. Gauhtier Real-time 3D ambisonics using Faust, processing, Pure Data, and OSC, in *Proc. of the 18 Int. DAFx Conference*, Trondheim, Norway, Nov 30 - Dec 3, 2015.

[9] P. Lecomte, P.-A. Gauthier, C. Langrenne, A. Berry, and A. Garcia A Fifty-Node Lebedev Grid And Its Applications To Ambisonics, Journal of the Audio Engineering Society, 64(11), (2016).

[10] M. Abramowitz and I. A. Stegun Handbook of mathematical functions, 10th Edition, Dover Publications, 1972.

[11] J. Daniel Spatial sound encoding including near field effect: Introducing distance coding filters and a viable, new Ambisonic format, in . *23rd AES conference*, Copenhagen, Denmark, 2003.

[12] Fons Adriaensen Near field filters for higher order Ambisonics, http://kokkinizita.linuxaudio.org/papers/hoafilt.pdf, 2006, Accessed: 2019-06-25.

doi:10.25836/sasp.2019.03

INVESTIGATION OF SWEET SPOT RADIUS OF SOUND RECONSTRUCTION SYSTEM BASED ON INVERSE FILTERING

Hiraku Okumura
Yamaha Corporation, Kyoto University
hiraku.okumura@music.yamaha.com

Makoto Otani
Kyoto University
otani@archi.kyoto-u.ac.jp

ABSTRACT

There are several methods aiming at sound field reconstruction a sound field, such as Higher-Order Ambisonics and Boundary Surface Control (BoSC) method. While the BoSC system aims to reconstruct a sound field within a volume surrounded by the boundary surface, some previous studies suggest that a reconstructed area, so-called a sweet spot, would be generated even outside of this controlled area. The authors investigated that the radius of a sweet spot for a BoSC system consisting of a spherical controlling surface and a loudspeaker array placed on a sphere. The results show that the radii of the microphone array do not affect the radii of the sweet spot, whereas the number of microphones could affect it. Furthermore, a simplified implementation only with sound pressure control could affect the radius of sweet spot.

1. INTRODUCTION

Sound field reconstruction techniques are very effective tools for a sound system of live-viewing or acoustical design in architecture. In a live-viewing system, listeners can enjoy highly realistic sound through the system. In addition, the system would allow acoustical designers to evaluate sound fields simulated in architectural spaces before their completion. There are several methods to realize sound field reconstruction, such as Higher-Order Ambisonics [1], Wave Field Synthesis [2], and Boundary Surface Control (BoSC) [3]. It is important to reconstruct a sound field within a broad region in order to allow a listener to look and move around, or to allow multiple listeners to experience the sound field at the same time, whereas reconstruction of sound field in a broad region necessitates a large number of microphones and loudspeakers. Fortunately, it has become easier and less expensive to handle a large number of devices thanks to the progress of computer technologies and network audio technologies, which contributes to a realization of broader reconstruction region, i.e. a broader sweet spot. In the BoSC system, both sound pressure and particle velocity on the boundary surface of a reconstruction region are controlled using inverse filters to reconstruct a sound field. While the BoSC system aims to reconstruct a sound field in a region surrounded by a boundary surface, some studies suggest that a sweet spot would be formed outside of this controlled region [4, 5]. However, the size of sweet spot in sound field reconstruction using inverse filtering is yet to be revealed. In this work, the authors numerically investigated the radius of sweet spot in sound field reconstruction with a spherical controlling surface and a spherical loudspeaker array.

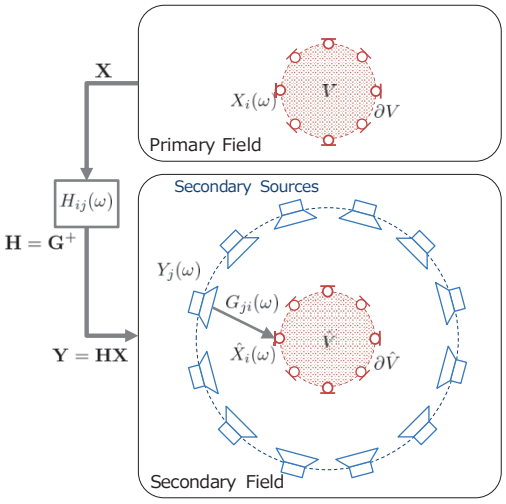


Figure 1. The concept of sound field reproduction using inverse filtering.

2. SOUND FIELD RECONSTRUCTION SYSTEM BASED ON INVERSE FILTERING

In this paper, the principle of boundary surface control [3] is employed as a theory for sound field reconstruction based on inverse filtering . Figure 1 illustrates the concept of sound field reconstruction by using inverse filtering. A sound field in a volume V in the primary field is reconstructed in a volume $\hat{V}$ in the secondary field by reconstructing both sound pressure and particle velocity on the boundary ∂V of the volume V at the corresponding position on the boundary $\partial\hat{V}$ of the volume $\hat{V}$ using secondary sources. It should be noted that particle velocity on the boundary $\partial\hat{V}$ is automatically reconstructed when sound pressure on the boundary $\partial\hat{V}$ is reconstructed, except at the frequencies associated with the internal Dirichlet problem of the volume $\hat{V}$. Therefore, generally, only sound

© Hiraku Okumura, Makoto Otani. Licensed under a Creative Commons Attribution 4.0 International License (CC BY 4.0). **Attribution:** Hiraku Okumura, Makoto Otani. "Investigation of Sweet Spot Radius of Sound Reconstruction System based on Inverse Filtering", 1st EAA Spatial Audio Signal Processing Symposium, Paris, France, 2019.

doi:10.25836/sasp.2019.16

pressures are reconstructed in this type of sound field reconstruction system.

Let $X_i(\omega)(i = 1, \cdots, Q)$ be the components of the angular frequency ω of the signals observed by Q microphones placed on the boundary ∂V in the primary field, and $\hat{X}_i(\omega)$ be the components of the angular frequency ω of the observed signals by Q microphones placed on the boundary $\partial \hat{V}$ in the secondary field. The input signals $Y_j(\omega)(j = 1, \cdots, M)$ to the M loudspeakers in the secondary field are designed to match $\hat{X}_i(\omega)$ to $X_i(\omega)$. $G_{ji}(\omega)$ is the transfer function between the j-th secondary loudspeaker and the i-th microphone in the secondary field. Using matrix representation, the system can be written as,

$$\mathbf{Y} = \mathbf{H}\mathbf{X} \tag{1}$$

$$\hat{\mathbf{X}} = \mathbf{G}\mathbf{Y} = \mathbf{G}\mathbf{H}\mathbf{X} \tag{2}$$

where,

$$\mathbf{Y} = [Y_1(\omega) \cdots Y_M(\omega)]^T \tag{3}$$

$$\mathbf{X} = [X_1(\omega) \cdots X_Q(\omega)]^T \tag{4}$$

$$\hat{\mathbf{X}} = \left[\hat{X}_1(\omega) \cdots \hat{X}_Q(\omega)\right]^T \tag{5}$$

$$\mathbf{H} = \begin{bmatrix} H_{1,1}(\omega) & \cdots & H_{1,Q}(\omega) \\ \vdots & \ddots & \vdots \\ H_{M,1}(\omega) & \cdots & H_{M,Q}(\omega) \end{bmatrix} \tag{6}$$

$$\mathbf{G} = \begin{bmatrix} G_{1,1}(\omega) & \cdots & G_{1,M}(\omega) \\ \vdots & \ddots & \vdots \\ G_{Q,1}(\omega) & \cdots & G_{Q,M}(\omega) \end{bmatrix}. \tag{7}$$

Finding $\mathbf{H}$ that satisfies $\mathbf{X} = \hat{\mathbf{X}}$ leads to

$$\mathbf{G}\mathbf{H} = \mathbf{I}. \tag{8}$$

In this paper, Moore-Penrose generalized inverse matrix $\mathbf{G}^+$ is used as $\mathbf{H}$,

$$\mathbf{H} = \mathbf{G}^+. \tag{9}$$

Kajita et al. [4] reported that a sweet spot could be larger than the boundary of the reconstructed region in the BoSC system. Further, Fazi et al. [5] suggested that sweet spot could be larger than boundary surface in the sound reconstruction based on an integral equation of the first kind.

3. SWEET SPOT

In this paper, the reconstruction performance of the system is evaluated by normalized reconstruction error (NRE) defined as

$$\mathrm{NRE}(\mathbf{x}) = \frac{|\hat{p}(\mathbf{x}) - p(\mathbf{x})|^2}{|p(\mathbf{x})|^2} \times 100\ [\%], \tag{10}$$

where $p(\mathbf{x})$ and $\hat{p}(\mathbf{x})$ are the sound pressure signals at the observation point $\mathbf{x}$ in the primary field and the secondary field respectively.

The sweet spot is defined as the area in which NRE is smaller than 4 % [6].

Kajita [4] defined the sweet spot as the area in which the S/N ratio is greater than 15 dB. This means that they defined the sweet spot as the region in which NRE is smaller than 3.16 %.

4. NUMERICAL SIMULATIONS

The numerical simulation assuming a free field was performed to investigate the size of the sweet spot in the sound field reconstruction based on inverse filtering. A 122-channels spherical loudspeaker array was assumed as secondary point sources, which located at the vertexes of a geodesic dome of 2.5-meter radius [7]. Further, a Q channels spherical microphone array was assumed as microphone capsules on a spherical surface, which located on a sphere with radius of R_V meters. The positions of microphone capsules were determined by the spherical Fibonacci spiral [7]. The center of the geodesic dome of the secondary sources was set to the origin, which also corresponds to the center of spherical microphone array.

A point source in the primary field was assumed at a random position whose distance to the origin was more than 2.5 meters. Transfer function between a point source $\mathbf{x}$ and a microphone capsule $\mathbf{y}$ is calculable as the free-field Green's function,

$$G(\omega) = \frac{e^{j\frac{\omega}{c}|\mathbf{x}-\mathbf{y}|}}{4\pi|\mathbf{x}-\mathbf{y}|}, \tag{11}$$

where ω is an angular frequency; c is the speed of sound; j is the imaginary unit. In the primary field, sound pressures $\mathbf{X}$ were calculated at the positions of the microphone capsules. In the secondary field, transfer functions $G_{ji}(\omega)(j = 1, \cdots, M, i = 1, \cdots, Q)$ between the $j-$th secondary source and the $i-$th microphone capsule were calculated to obtain $\mathbf{H} = \mathbf{G}^+$. Finally, the input signal $\mathbf{Y}$ to the secondary sources were derived as $\mathbf{Y} = \mathbf{H}\mathbf{X}$. In this section, sound pressures were calculated in the frequency domain and its frequency interval was 7.8125 Hz.

To evaluate the sweet spot size, sound pressures were calculated inside V in the primary field and $\hat{V}$ in the secondary field at the grid points of 5.0 cm intervals in x and y-directions in a square with a 2.5 meters sides and centered on the origin.

Figure 2 depicts NRE [%] for the point source at $(-8, 0, 0)$ in the primary field. The group of gray-color points drawn around the origin in Figure 2 represents the microphone capsules, that is a 64-channels microphone array of $R_V = 0.05$ m. The radius of the sweet spot is defined as the minimum distances between the origin and the point where NRE is smaller than 4 %.

Figure 3 demonstrates the radii of the sweet spot for a 64 channels spherical microphone array of radius $R_V = 0.05$ m. The gray lines indicate the radii of the sweet spot of the sound field reconstruction system for point sources placed at the randomly generated positions in the primary field. The thick line represents the average of these radii of sweet spot. In the following sections, the radius of sweet spot is defined as the average of the radii of the sweet spot among 100 randomly generated point sources.

doi:10.25836/sasp.2019.16

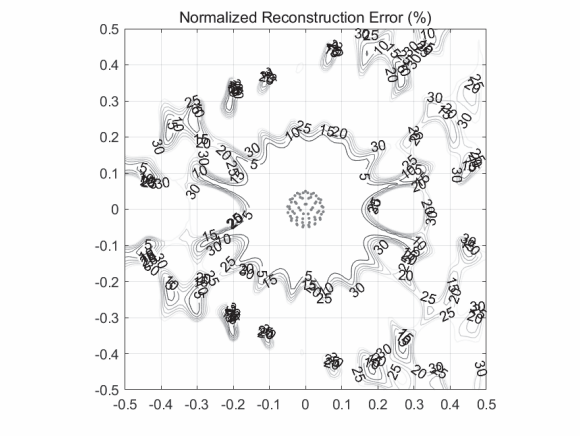


Figure 2. NRE (%) at 2kHz for spherical microphone array of $R_V = 0.05$ m, $Q = 64$.

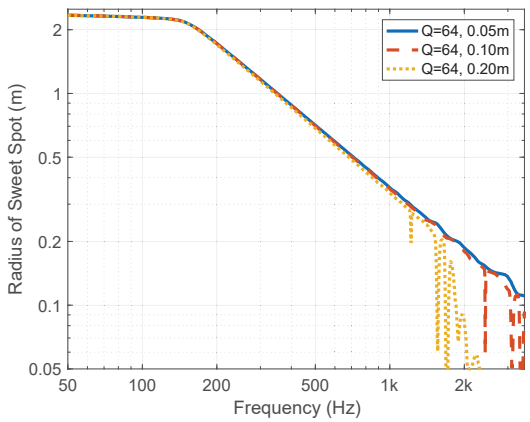


Figure 4. Radii of sweet spot for 64-channel spherical microphone arrays of various R_V.

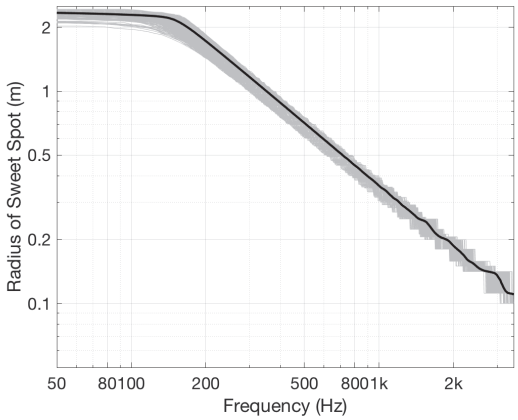


Figure 3. Radii of the sweet spot for the spherical microphone array of $R_V = 0.05$ m, $Q = 64$. The thick line shows the average of these radii.

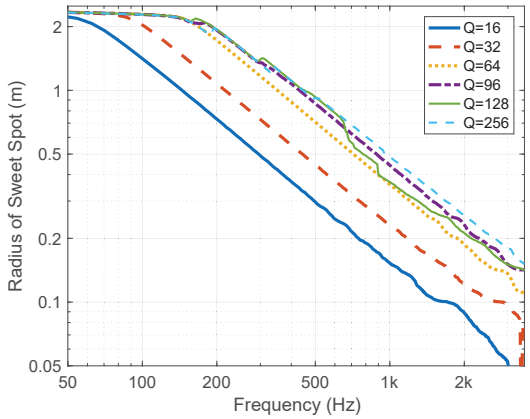


Figure 5. Radii of sweet spot for the sperical microphone array of $R_V = 0.05$ m with various Q.

4.1 Effects of radius and number of microphone capsules on sweet spot radius

Figure 4 illustrates radii of the sweet spot for 64-channel spherical microphone arrays of $R_V = 0.05$ m, 0.10 m and 0.20 m. The radii of the sweet spot do not differ among the three radii of the microphone array at frequencies below 1.1 kHz.

When $R_V = 0.1$ m and 0.2 m, the radii of the sweet spot drop suddenly at around 2.4 kHz and 1.2 kHz respectively. The minimum microphone intervals of the spherical microphone array of $R_V = 0.1$ m and 0.2 m are 0.0386 m and 0.0772 m, respectively. These lengths correspond to a quarter of the wavelength λ for 2.4 kHz ($\lambda = 0.0354$ m) and 1.2 kHz ($\lambda = 0.0773$ m). Therefore, these frequencies could be regarded as the upper limit frequencies for the sound reconstruction system using the spherical microphone array of $R_V = 0.1$ m and 0.2 m.

Figure 5 depicts the radii of the sweet spot for the spherical microphone array of radius $R_V = 0.05$ m with number of the microphone capsules $Q = 16, 32, 64, 96, 128$ and 256. Generally, the radius of the sweet spot increases as the number of the microphone capsules Q increases. However, the radii of the sweet spot do not increase prominently when $Q > 96$. In this study, the number of secondary sources is 122, and the system defined as equation (8) is overdetermined when $Q = 128$ and 256. Therefore, the sound reconstruction system does not have an exact solution and this is considered to be the reason why the radii of the sweet spot do not increase when $Q > 96$. Thus, when the interval between microphone capsules of the spherical microphone array is greater than a quarter of the wavelength, the radius of the sweet spot depends on the number of microphone capsules of the spherical microphone array and does not depend on its radius [8].

Fazi [5] formulated a sound field reproduction system that reconstructs only the sound pressure on the boundary with an integral equation of the first kind, and suggested the correspondence of the Higher-Order Ambisonics (HOA) by expressing it using spherical harmonics. As described above, usually only sound pressures are reconstructed in the BoSC system. This corresponds to the method proposed by Fazi [5]. It is known that the size of the sweet spot of HOA depends on the number of divisions of the surface of the spherical microphone array

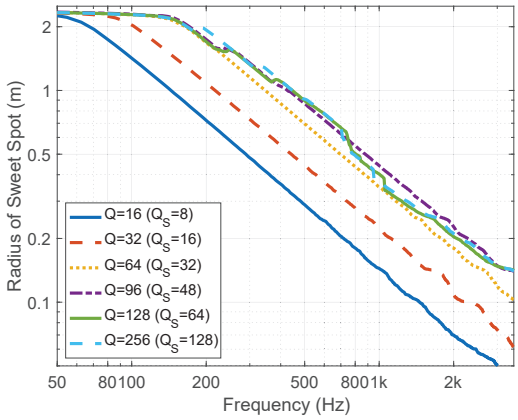


Figure 6. Radii of sweet spot for double-layered spherical microphone array of $R_V = 0.05$ m with a variety of Q.

(corresponding to the number of microphones). Therefore, the size of sweet spot in a sound field reconstruction that reconstructs only the sound pressure on the boundary can be regarded similar to one in HOA, which corresponds well with the above results in the current paper.

4.2 Effects of double-layered microphone arrays on the radii of sweet spots

Based on the BoSC principle, in order to reconstruct both the sound pressure and the particle velocity on the boundary, a double-layered microphone array has been proposed with the microphone capsules arranged at positions with offset both inward and outward from the boundary [3]. Assuming that the boundary surface is discretized by Q_S control points, a double-layered microphone array uses twice as many microphones as a single-layer microphone array that is used to reconstruct only the sound pressure on the boundary. Thus, $Q = 2Q_S$ for a double-layered microphone array whereas $Q = Q_S$ for a single-layer one.

Figure 6 shows the radii of the sweet spot for a spherical double-layered microphone array of radius $R_V = 0.05$ m with $Q = 16, 32, 64, 96, 128$ and 256, where the offset of inner and outer layers of the microphone capsule from the boundary is 2.5 mm. It is observed that the radii of the sweet spot in the sound field reconstruction using a double-layered spherical microphone array are very similar to that using a single-layer spherical microphone array in Figure 5. However, considering the number of divisions of the surface of the spherical microphone array, $Q_S = Q/2 = 8, 16, 32, 48, 64$, and 128, it appears that for the same number of divisions of the surface of the spherical microphone array, the size of the sweet spot of the sound field reconstruction system can be made greater by using a double-layered spherical microphone array. This indicates that a microphone array implemented using a less-dense array of microphone capsules can yield a larger sweet spot.

5. CONCLUSION

The size of the sweet spot in a sound reconstruction system using inverse filtering was investigated.

A sweet spot was generated not only inside but also outside of the controlled region by reconstructing sound pressures on the boundary of the volume. The numerical results revealed that, when the interval between microphone capsules of the spherical microphone array is greater than a quarter of the wavelength, the size of sweet spot in the sound field reconstruction depends on the number of microphone capsules in the spherical microphone array, and it does not depend on the radius of the spherical microphone array. Furthermore, the results also show that the size of sweet spot in the sound field reconstruction can be expanded by using a double-layered spherical microphone array to reconstruct both sound pressures and particle velocities on the boundary of the volume. This suggests that a larger sweet spot can be generated even when using a sparse (less dense) microphone array.

6. REFERENCES

[1] J. S. Bamford and J. Vanderkooy, "Ambisonic sound for us.," in *Proc. of Audio Eng. Soc. Conv.*, (New York, U.S.A.), 1995.

[2] A. J. Berkhout, "Ambisonic sound for us.," *Journal of Audio Eng. Soc.*, vol. 36, no. 12, pp. 977–995, 1988.

[3] S. Ise, "A principle of sound field control based on the kirchhof-helmholtz integral equation and the theory of inverse systems.," *Acoustica - Acta Acoustica*, vol. 85, pp. 78–87, 1999.

[4] Y. Kajita, K. Miura, Y. Watanabe, and S. Ise, "A study of relationship between microphone-positions and bosc-system reproduced-area using sound field estimation-method with lu decomposition.," in *Proc. of Research Meeting of the Acoustical Society fo Japan (in Japanese)*, (Tokyo, Japan), pp. 493–494, 2018.

[5] F. M. Fazi, P. A. Nelson, J. E. N. Christensen, and J. Seo, "Surround system based on three dimensional sound field reconstruction.," in *Proc. of Audio Eng. Soc.*, (San Francisco, USA), 2008.

[6] D. B. Ward and T. D. Abhayapala, "Reproduction of a plane-wave sound field using an array of loudspeakers.," *IEEE Trans. on Speech and Audio Processing*, vol. 9, no. 6, pp. 697–707, 2001.

[7] S. Kaneko, T. Suenaga, and H. Okumura, "Development of a 64-channel spherical microphone array and a 122-channel loudspeaker array system for 3d sound field capturing and reproduction technology research.," in *Proc. of Audio Eng. Soc. Conv.*, (Milan, Italy), 2018.

[8] H. Okumura and M. Otani, "Investigation of the sweet spot of sound field reproduction using inverse filtering.," in *Proc. of the Architectual Acoustics Research Meeting (in Japanese)*, (Sapporo, Japan), 2018.

doi:10.25836/sasp.2019.16

DETECTING THE DIRECTION OF EMERGENCY VEHICLE SIRENS WITH MICROPHONES

Noam Shabtai
User Experience Technologies
GM Advanced Technical Center, Israel
shabtai.noam@gmail.com

Eli Tzirkel
User Experience Technologies
GM Advanced Technical Center, Israel
eli.tzirkel@gm.com

ABSTRACT

As drivers we use both our eyes and ears as sensors whereas current autonomous vehicle sensors and decision making do not rely on sound. Sound is particularly important in cases involving Emergency Vehicle (EV) sirens, horn, shouts, accident noise, a vehicle approaching from a sharp corner, poor visibility, and other instances where there is no direct line of sight or it is limited. In this work the Direction of Arrival (DoA) of an EV is detected using microphone arrays. The decision of an Autonomic Vehicle (AV) whether to yield to the EV is then dependent on the estimated DoA.

1. INTRODUCTION

As human drivers, we are capable of using both our eyes and ears to get useful information in a traffic environment [1]. Hence, the development of AVs is challenging in that the vehicle should be able to perform no worse than a human driver (if not better), and be able to collect data from the external environment under the same conditions [2]. Rapidly moving objects such as other vehicles or bicycles, slower objects such as pedestrians, and static objects such as parked cars and barriers should be all sensed by the AV and used in algorithms for correct decision making [3]. Several sensors can be used for sensing these objects; e.g., radar [4], cameras [5], and microphones [6]. In cases where an object that emits sound is too far away or near but concealed from the car, sound recorded by microphones may be the only reliable source of information. There are vast numbers of cases where sound information is important, including EV sirens, horn, shouts, accident noise, a vehicle approaching from a sharp corner, and poor visibility.

Recently, Waymo shared a report with the US Transport Department where microphones are used as "supplemental sensors" [7]. Furthermore, Waymo has developed microphones that let its robocars hear sounds twice as far away as previous sensors while also letting them discern where the sound is coming from [8]. Moreover, a video is available on the web, where it is shown how Waymo is learning to recognize emergency vehicles in Arizona, using sound and light [9].

In this work, we focus on the DoA estimation of EVs using microphone arrays. The estimated DoA can be used to decide whether to yield to an approaching EV. In practice, an EV siren is detected prior to the estimation of its DoA; however, this is a different and easier problem and can be handled using audio signature, and therefore is not addressed in this work. The DoA is estimated using a Multiple Signal Classification (MUSIC)-based algorithm and includes time smoothing technique to improve the reliability of the estimated DoA values. For the DoA estimation using internal microphones we implement a transfer function projection. Here, the DoA can be roughly estimated to determine whether the EV is approaching from behind, in which case the decision of the AV should be to yield to the EV.

Both internal and external microphone array approaches were investigated for their performance. The rational for using an external microphone array is that the results are more reliable and free-field steering vectors can be used; however, the microphones need to be protected from wind. Internal microphones have the advantage of already being available in the car for other applications; e.g., beamforming for the enhancement of Automatic Speech Recognition (ASR) performance. Unfortunately, free-field steering vectors cannot be used and transfer functions were measured with a lower spatial resolution instead. The results showed that despite the additional cost of mounting an external microphone array, it is recommended since the estimated DoA values are far more reliable than the ones achieved using the internal array.

2. DOA ESTIMATION

In this work a MUSIC-based algorithm was used for DoA estimation. Let $s(t,f)$ be the source signal in the Short Time Fourier Transform (STFT) domain. This signal is then received at the m'th microphone as

$$x_m(t,f) = a_m(f,\theta_i)\, s(t,f), \tag{1}$$

where $a_m(\cdot,\theta_i)$ is the transfer function from a source at direction θ_i to the m'th microphone. The signal vector at

© Noam Shabtai, Eli Tzirkel. Licensed under a Creative Commons Attribution 4.0 International License (CC BY 4.0). **Attribution:** Noam Shabtai, Eli Tzirkel. "Detecting the Direction of Emergency Vehicle Sirens with Microphones", 1st EAA Spatial Audio Signal Processing Symposium, Paris, France, 2019.

doi:10.25836/sasp.2019.22

all M microphones can be represented as

$$\begin{aligned}\mathbf{x}(t,f) &= [x_1(t,f), x_2(t,f), \ldots, x_M(t,f)]^T \\ &= \mathbf{a}(t,f)\, s(t,f), \end{aligned} \quad (2)$$

where

$$\mathbf{a}(f,\theta_i) = [a_1(f,\theta_i), a_2(f,\theta_i), \ldots, a_M(f,\theta_i)]^T \quad (3)$$

is referred to as the *steering vector* from direction θ_i at frequency f.

Practically, the signal vector $\mathbf{x}$ is received by the microphones and used to calculate $\hat{\theta}_i$, the estimation of θ_i. The autocorrelation of $\mathbf{x}$ is given by

$$\mathbf{R}_{\mathbf{x}}(t,f) = E\left[\mathbf{x}(t,f)\,\mathbf{x}^H(t,f)\right]. \quad (4)$$

Assuming full rank of $\mathbf{R}_{\mathbf{x}}$, it has M eigenvectors. The eigenvector with the largest eigenvalue is associated with the signal space, and all the rest are associated with the noise space. In general, the MUSIC algorithm is designed for any number of sources up to $M-1$, but in this application only one source was of interest. Hence, if the eigenvectors $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_M$ are sorted in descending order, the noise space eigenmatrix is defined by

$$\tilde{\mathbf{U}}(t,f) = [\mathbf{u}_2(t,f), \mathbf{u}_3(t,f), \ldots, \mathbf{u}_M(t,f)]^T. \quad (5)$$

The MUSIC spectrum P is then calculated using the noise-space eigenmatrix $\tilde{\mathbf{U}}$ and the steering vector of the hypothetical DoA $\mathbf{a}(t,f,\theta_h)$ to form

$$P(t,f,\theta_h) = \frac{\|\mathbf{a}(f,\theta_h)\|^2}{\mathbf{a}(f,\theta_h)^H \tilde{\mathbf{U}}(t,f)\,\tilde{\mathbf{U}}(t,f)^H \mathbf{a}(f,\theta_h)}. \quad (6)$$

As a first step, the frequency for which the MUSIC spectrum is calculated is selected as the one with highest energy in the received signal at the first microphone. That is

$$f_0(t) = \arg\max_f |x_1(t,f)|, \quad (7)$$

and then the estimated DoA is given by

$$\hat{\theta}_i(t) = \arg\max_{\theta_h} P(t, f_0(t), \theta_h). \quad (8)$$

Temporal smoothing is performed to prevent the consideration of non-realistic estimated DoA values. If in Eq. (8) the raw estimated DoA value $\hat{\theta}_i(t)$ is given using the plain maximum value of the MUSIC spectrum $P(t,f,\theta_h)$, then the smoothed DoA is

$$\hat{\theta}_s(t) = \arg\max_{\theta_h} \int_{t-T}^{t} P(\tau, f_0(\tau), \theta_h)\, d\tau. \quad (9)$$

Frequency smoothing can be used to select frequencies f_0 that are near the previously selected frequency since the siren signal is essentially an ascending and decreasing chirp signal. However based on some preliminary results, it was decided not to use frequency smoothing.

3. EXTERNAL MICROPHONE ARRAY

3.1 Hardware

The external microphone array consisted of 4 Micro-electromechanical system (MEMS) microphones selected from 32, as can be seen in Fig. 1, arranged as a square of dimensions $5 \times 3\text{cm}^2$. The grid dimensions of the microphone array was taken into consideration when calculating the free-field steering vectors for the DoA estimation algorithm. The external microphone array was placed outside the car and mounted on the roof as can be seen in Fig. 2.

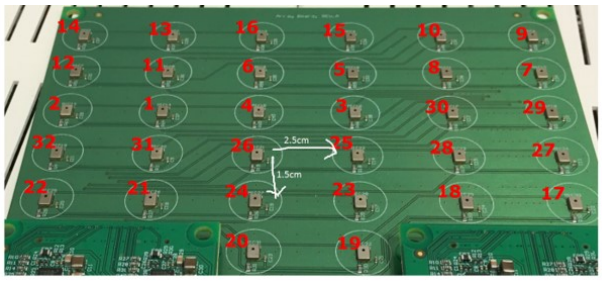


Figure 1: External microphone array.



Figure 2: External microphone array on the roof of the car.

3.2 Steering Vector

The advantage of the external microphone array is that for the DoA estimation algorithm, the steering vectors can be roughly considered like those in free field. Since the EV is far away, the incident wave form can be considered to be a plane wave, as shown in Fig. 3, where θ_i is the DoA angle, and r_m, θ_m are the distance and angle of the m'th microphone from the origin of the microphone array, respectively.

The free-field steering vector can therefore be calculated in an x-y plane. Let f be the frequency of the sound that is generated by the EV. At this frequency the wave number is $k = \frac{2\pi f}{c}$ where $c = 343\frac{\text{m}}{\text{s}}$ is the speed of sound. The frequency response from the source at θ_i to the m'th microphone with regard to the origin is given by

$$a_m(f,\theta_i) = e^{-jkr_m\cos(\theta_i-\theta_m)}, \quad (10)$$

neglecting differences in amplitude attenuation from the source to the origin and to the microphones. The steering vector of the array that contains M microphones is given

doi:10.25836/sasp.2019.22

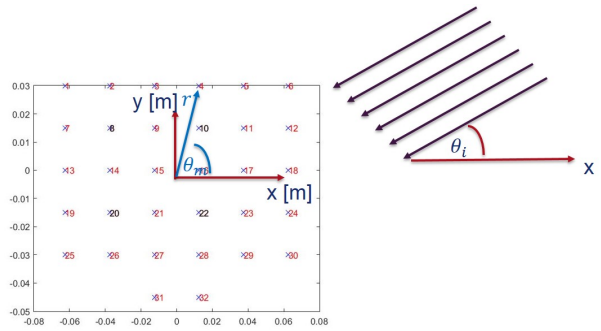


Figure 3: Plane wave propagating to the external microphone array.

by

$$\mathbf{a}\left(f,\theta_i\right)=\left[a_1\left(f,\theta_i\right),a_2\left(f,\theta_i\right),\ldots,a_M\left(f,\theta_i\right)\right]^T. \tag{11}$$

3.3 EV Experimental Results

The external microphone array was mounted on the roof of the XTS car. The car was parked near a hospital. The EVs were ambulance vehicles recorded arriving to or departing from the hospital. The parked car and the EV station can be seen in Fig. 4.

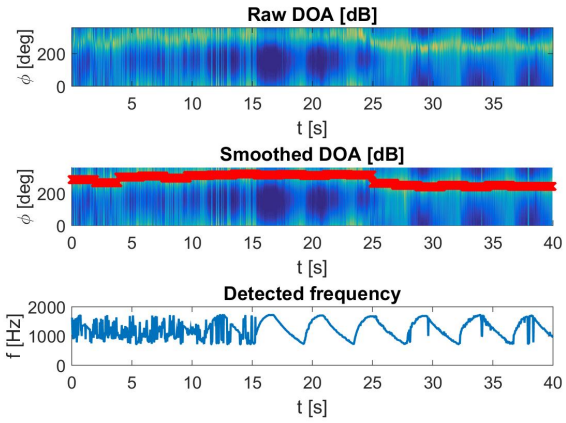


Figure 4: XTS car parked near an EV station.

The case where an EV approached from the opposite lane and made its way to the hospital is shown for example. Basically, at first the DoA comes from the frontal direction, and then switches to behind the car. Figure 5 shows the MUSIC spectrum, the estimated DoA, and the selected frequency for this case. At $t = 25$s the peak of the MUSIC spectrum shifted from values near 360° to values near 180°. It has been detected that the porches of the microphone array board reflect the sound, and therefore even though the EV was on the left side, the peak values appeared at angles that corresponded to the right side. Nevertheless, it was easy to determine when the EV was in front or behind the car. In this case, the decision of an AV should be to continue normal driving and not to yield to the EV.



(a)



(b)

Figure 5: (a) An EV is approaching from the frontal direction (b) MUSIC spectrum and estimated DoA show switching from frontal ($\sim 360°$) to back ($\sim 180°$) direction. The selected frequency matches the siren.

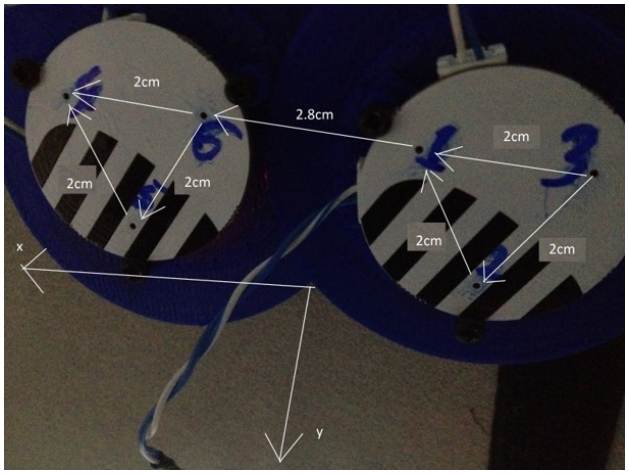
4. INTERNAL MICROPHONE ARRAY

4.1 Hardware

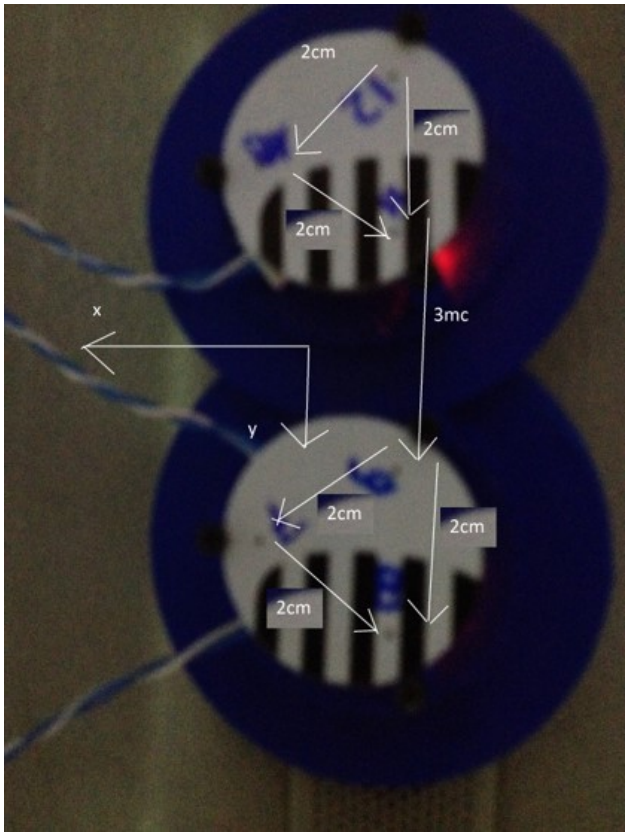
The internal microphone array consisted of different microphones but the same dedicated hardware for sound acquisition as for the external microphone array. The array contained two sub arrays with 3 MEMS each, together forming an array of 6 microphones, from which a new subset of microphone could be selected to form a different microphone array configuration.

The internal microphone array was placed inside the car and mounted either above the rear-right or the front-left passenger, corresponding to the placement of arrays for speech recognition or hands-free calls. A subset of 4 microphones can be selected to form an end-fire configuration above the rear-right passenger as presented in Fig. 6a, or a broad-side configuration above the front-left passenger, as can be seen in Fig. 6b. For the rear-right array, the distance between any pair of microphones on each sub array was 2cm, and the minimum distance between microphones from different sub arrays was 2.8cm, as shown in Fig. 6a. For the front-left array, the distance between any pair of microphones on each sub array was 2cm, and the minimum distance between microphones from differ-

ent sub arrays was 3cm, as shown in Fig. 6b.



(a)



(b)

Figure 6: Distance between microphones in the internal array, a subset of 4 microphones forms (a) an end-fire array above the rear-right passenger and (b) a broad-side array above the front-left passenger.

4.2 Steering Vector

In the case of the internal microphone array, the steering vector cannot be calculated using a free-field representation and instead, the frequency response of the car from each DoA needs to be considered. Therefore, rather than an analytic calculation of the steering vector with high spatial resolution as used by the external array, in the case of the internal array the transfer function needs to be measured in a quiet area with much lower spatial resolution. Since it is very difficult to measure the Acoustic Transfer Function (ATF) from the source to each microphone, the Relative Transfer Function (RTF) was used instead, in such a way that at each microphone the frequency response was calculated relative to the ATF at the 1st microphone.

Let h_m be the ATF from the source to the m'th microphone. The RTF is given by

$$RTF_m(f) = \frac{h_m(f)}{h_1(f)}. \tag{12}$$

If an acoustic source emits a signal $x(f)$ and assuming a noise signal $n_m(f)$ at the m'th microphone, the recorded signal at the m'th microphone is

$$\begin{aligned} y_m(f) &= h_m(f)x(f) + n_m(f) \\ &= RTF_m(f)h_1(f)x(f) + n_m(f) \\ &= RTF_m(f)\left(y_1 - n_1(f)\right) + n_m(f) \\ &= RTF_m(f)y_1(f) + \tilde{n}_m(f), \end{aligned} \tag{13}$$

where $\tilde{n}_m(f) = n_m(f) - RTF_m(f)n_1(f)$.

4.3 Wienner Filter

The estimation of the RTF based on the Wiener filter solution that minimizes the variance of the error is performed using

$$\widehat{RTF}_m(f) = \arg\min_{RTF} E\left[\left|y_m(f) - RTFy_1(f)\right|^2\right]. \tag{14}$$

which leads to

$$\widehat{\mathbf{RTF}}_m = \left[\frac{y_1^\star(1)}{\|y_1(1)\|^2}y_m(1), \ldots, \frac{y_1^\star(F)}{\|y_1(F)\|^2}y_m(F)\right]^T. \tag{15}$$

.

4.4 Generalized Eigenvalue Decomposition (GEVD)

Defining vectors with microphone indices rather than frequencies as coordinates

$$\mathbf{y}(f) \triangleq [y_1(f), y_2(f), ..., y_M(f)]^T \tag{16}$$
$$\mathbf{n}(f) \triangleq [n_1(f), n_2(f), ..., n_M(f)]^T \tag{17}$$
$$\mathbf{h}(f) \triangleq [h_1(f), h_2(f), ..., h_M(f)]^T \tag{18}$$

yields the following vector form

$$\mathbf{y}(f) = \mathbf{h}(f)x(f) + \mathbf{n}(f) \tag{19}$$

to Eq. (13). Applying the autocorrelation operator to Eq. (19) yields

$$\mathbf{R}_\mathbf{y}(f) = \sigma_x^2(f)\mathbf{h}(f)\mathbf{h}^H(f) + \mathbf{R}_\mathbf{n}(f). \tag{20}$$

The process of GEVD of $\mathbf{R}_\mathbf{y}(f)$ with respect to $\mathbf{R}_\mathbf{n}(f)$ relates the generalized eigenvalues $\lambda_m(f)$ to the corresponding generalized eigenvectors $\mathbf{v}_m(f)$ by solving

$$\mathbf{R}_\mathbf{y}(f)\mathbf{v}_i(f) = \lambda_i(f)\mathbf{R}_\mathbf{n}(f)\mathbf{v}_i(f), \qquad i = 1, 2, \ldots, M \quad , \tag{21}$$

doi:10.25836/sasp.2019.22

assuming that the rank and the number of microphones are identical and equal to M.

Assuming that the eigenvectors are sorted in descending order

$$\lambda_1(f) \geq \lambda_2(f) \geq \cdots \geq \lambda_M(f) \quad , \tag{22}$$

The generalized eigenvector that corresponds to the largest generalized eigenvalue is a rotated and scaled form of the ATF [10]. The RTF can be calculated using

$$\widehat{\mathbf{RTF}}_i(f) = \frac{\mathbf{R_n}(f)\mathbf{v}_1(f)}{(\mathbf{R_n}(f)\mathbf{v}_1(f))_{(1)}}, \tag{23}$$

where subscript $(\cdot)_{(1)}$ indicates the first coordinate of a vector.

4.5 RTF Estimation Performance

The estimation of the RTF was evaluated using Signal to Distortion Ratio (SDR). The SDR was used to calculate the distortion between the signal recorded by a microphone in the array y_m to the signal that is generated by filtering the signal recorded from the first microphone y_1 with RTF_m:

$$SDR_m(f_1, f_2) = \frac{1}{f_2 - f_1} \int_{f_1}^{f_2} \frac{\|y_m(f)\|^2}{\left\|y_m(f) - y_1(f)\widehat{RTF}_m(f)\right\|^2} df \tag{24}$$

The SDR values are displayed in Fig. 7 and Fig. 8 for the performance evaluation of the RTF estimation process using the internal microphone array in the broad-side and end-fire configurations, respectively. The RTF was evaluated using a controlled measurement where the recording car was placed in an isolated parking spot, and another car displayed a sweep signal using a speaker mounted on its roof from different directions with a resolution of $45°$. The angle of direction is displayed on the horizontal axes, and the microphone index m is displayed on the vertical axes. The corresponding SDR value is expressed in dB units using gray levels.

For the case examined in this work, the most interesting directions are $0°$ and $180°$, which correspond to the frontal and back directions, respectively. For these directions, the RTF was estimated better for the broad-side configuration than for the end-fire configuration. Comparing Fig. 7a to Fig. 7b, and also comparing Fig. 8a to Fig. 8b, shows that the RTFs were estimated better using the LS method than when using the GEVD method for all directions and all microphones.

4.6 EV Experimental Results

Only front and back RTFs were used as steering vectors. Figure 9 shows the MUSIC spectrum and DoA estimation results for the case where an EV is approaching the car from the opposite lane. The results in the figures show that using the end-fire array it is impossible to determine whether the EV was behind or in front of the car. The DoA was estimated better using the broad-side array. This result may appear surprising, since one would expect that the

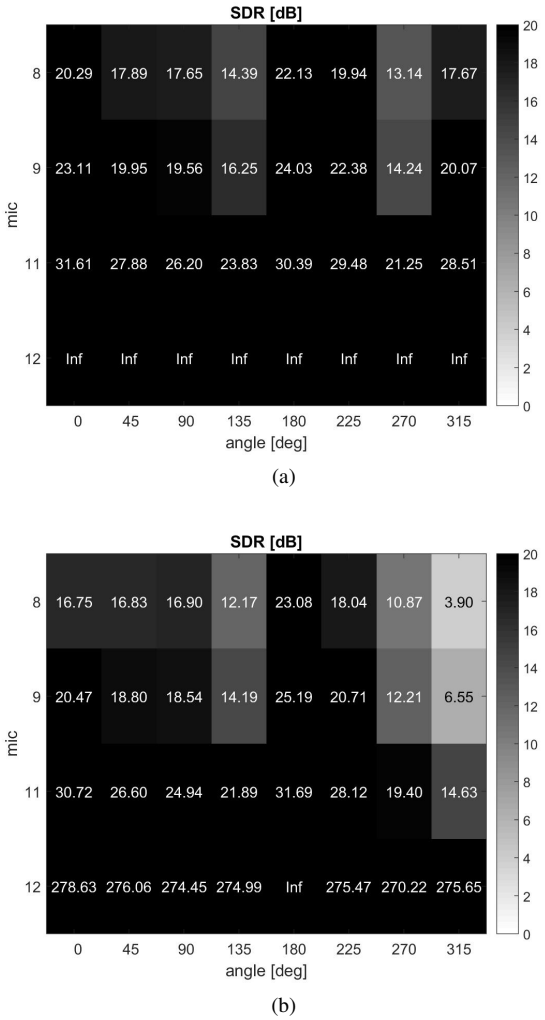


Figure 7: Evaluation of the RTF estimation using SDR for the internal array in the broad-side configuration using the (a) LS and (b) GEVD estimation methods.

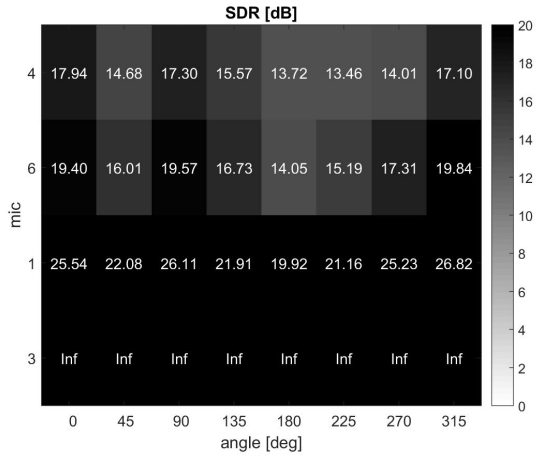
symmetry of the broad-side array around the driving direction would have caused an ambiguity for waves approaching from the front or from the back. However, as explained in the previous section, the RTFs are estimated using each microphone with less distortion using the broad-side array than using the end-fire array. Regardless, the difficulty of estimating the DoA and the lower angular resolution was greater in the case of the internal microphone array than in the case of the external one.

5. CONCLUSION

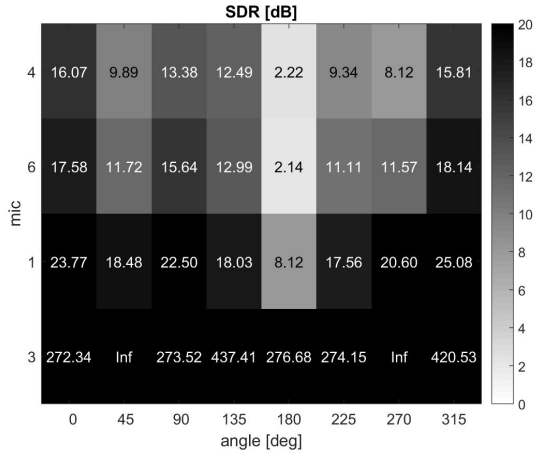
The feasibility of detecting the direction of an approaching EV was validated using an external microphone array equipped with 4 microphones. An algorithm for using internal microphones was developed in order but found to be inferior to an external array.

6. REFERENCES

[1] M. Sivak, "The information that driver use: is it indeed 90% visual?," *Perception*, vol. 25, no. 9, pp. 1081–

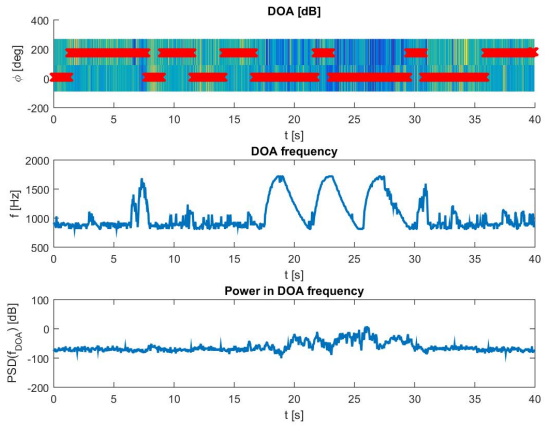


(a)

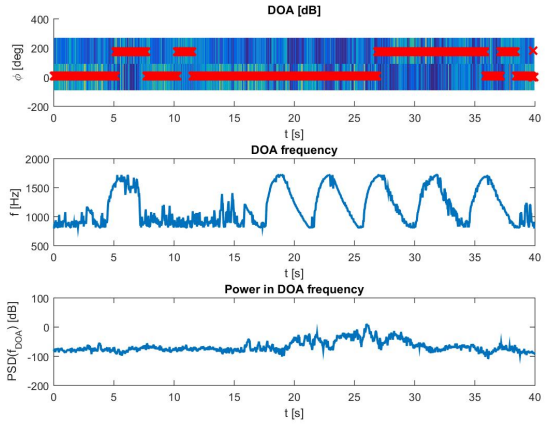


(b)

Figure 8: Fig. 7 repeated for end-fire configuration.



(a)



(b)

Figure 9: MUSIC spectrum and estimated DoA for an EV approaching from the frontal ($\sim 360°$) to back ($\sim 180°$) direction. The selected frequency matches the siren.

1089, 1996.

[2] Y. W. Seo and C. Urmson, "A perception mechanism for supporting autonomous intersection handling in urban driving," in *Proc. IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 1830–1835, 2008.

[3] A. Schaub, D. Baumgartner, and . Burschka, "Reactive obstacle avoidance for highly maneuverable vehicles based on a two-stage optical flow clustering," *IEEE Transactions on Intelligent Transportation Systems*, vol. 18, no. 8, pp. 2137–2152, 2016.

[4] I. Ruiz, D. Aufderheide, and U. Witkowski, "Radar sensor implementation into a small autonomous vehicle," in *Advances in Autonomous Mini Robots* (U. Rückert, S. Joaquin, and W. F. W., eds.), Berlin, Heidelberg: Springer, 2012.

[5] M. Pereira, D. Silva, V. Santos, and P. Dias, "Self calibration of multiple lidars and cameras on autonomous vehicles," *Robotics and Autonomous Systems*, vol. 83, pp. 326–337, 2016.

[6] D. I. Ferguson and J. Zhu, "Controlling autonomous vehicle using audio data," Mar. 2014. US Patent 8,676,427 B1.

[7] S. Collie, "Waymo's driverless cars make thousands of decisions every second to keep you alive," 2017. https://www.caradvice.com.au/591773/waymo-safety-report-sheds-light-on-the-life-of-a-self-driving-car/.

[8] J. Stewart, "Driverless cars need ears as well as eyes," 2017. https://www.wired.com/story/driverless-cars-need-ears-as-well-as-eyes/.

[9] M. McFarland, "Waymo's self-driving vans learn how to drive near police cars," 2017. https://money.cnn.com/2017/06/30/technology/future/waymo-chandler-arizona/index.html.

[10] S. Markovich, S. Gannot, and I. Cohen, "Multichannel eigenspace beamforming in a reverberant noisy environment with multiple interfering speech signals," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, pp. 1071–1086, Aug 2009.

doi:10.25836/sasp.2019.22

ACOUSTIC SIMULATION OF BACH'S PERFORMING FORCES IN THE THOMASKIRCHE

Braxton Boren[1] **Daniel Abraham**[1] **Rogerio Naressi**[1]
Elizabeth Grzyb[1] **Brandie Lane**[2] **Daniel Merceruio**[3]
[1] Department of Performing Arts, American University, Washington, DC, USA
[2] US Military Academy, West Point, USA
[3] Independent Producer, North Carolina, USA
boren@american.edu

ABSTRACT

Since room acoustics profoundly affect musicians' performance style, renderings of early music should account for historical acoustics as well as instrument design and ensemble size. This paper describes the the recording process for the project *Hearing Bach's Music as Bach Heard It*, which uses acoustic measurements and geometric acoustic modeling to render the soundscape of Bach's Thomaskirche in Leipzig, Germany in 1539 and 1723, the year Bach arrived. This historical model was used to calculate section-to-section binaural impulse responses for each section of musicians on a Bach cantata. Using real-time convolution and close-miking, the musicians were recorded while performing in the virtual church at both time periods, with audible differences between the two. Some discussion follows as to the optimal method for arranging dry multitrack recordings of historical works when separate anechoic chambers are not available for each musician.

1. INTRODUCTION

The study of cultural heritage sites is often undervalued when the objects of their study are well-preserved: indeed the very preservation of historically significant sites often leads us to take them for granted, assuming they will always be with us. This cultural tendency was briefly arrested after the colossal fire at Notre Dame de Paris on 15 April 2019. Because of the specific musicological importance of Notre Dame's reverberant acoustic, some feared that the church's soundscape might be permanently lost [1]. However, due to the extensive measurement and simulation work of Postma, Katz, and others, the historical acoustics can be rebuilt including the contributions of individual surfaces and materials [2]. One hopes that this tragedy will give greater urgency to the task of acoustically cataloging historically significant sites which have not been as extensively studied as Notre Dame.

Besides applications in gaming and virtual reality, acoustic simulation technology also has potential to transform historical musicology by rendering audible lost soundscapes from the past [3–8]. This in turn may allow the role of historical spaces to be explored along with historical instruments and ensembles [9]. This requires individual recordings of each instrument being simulated, which involves many performers in the case of much of the Western classical repertoire. Simulation is also complicated by the fact that trained performers naturally alter their tempo and articulation in different acoustic environments [10].

Another era in European classical music that is distinctly tied to a specific place is the mature compositional period of J.S. Bach (1685–1750) in Leipzig, Germany. This paper describes the project *Hearing Bach's Music as Bach Heard It*, which addresses the acoustic sound field presented by the Thomaskirche in Leipzig, Germany in 1723, the start of J.S. Bach's tenure as cantor. In addition, alterations to the virtual model also allow an investigation of the acoustics before the alterations to the church by the Lutheran Reformers in the 16th century. An entire Bach cantata, *Herz und Mund und Tat und Leben* (BWV 147), has been recorded to virtually place the musicians in the virtual church at these two points in the space's history.

2. BACKGROUND

2.1 Bach and Acoustics

Much of the historically informed performance movement has focused on the work of Bach, including the construction of historical instruments more similar to what Bach's musicians would have used [11]. In addition, musicologists disagree over whether Bach used a small (1 voice per part) or large (3-4 voices per part) chorus for his cantatas in the Leipzig Thomaskirche and other churches [12–15]. One recent work arguing for a much larger vocal ensemble relies almost exclusively on general acoustical arguments but produces no quantitative evidence in support of this conclusion [16].

© Braxton Boren, Daniel Abraham, Rogerio Naressi, Elizabeth Grzyb, Brandie Lane, Daniel Merceruio. Licensed under a Creative Commons Attribution 4.0 International License (CC BY 4.0). **Attribution:** Braxton Boren, Daniel Abraham, Rogerio Naressi, Elizabeth Grzyb, Brandie Lane, Daniel Merceruio. "Acoustic Simulation of Bach's Performing Forces
in the Thomaskirche", 1st EAA Spatial Audio Signal Processing Symposium, Paris, France, 2019.

doi:10.25836/sasp.2019.39

Recordings of Bach's choral works tend to ignore the issue of room acoustics, or implicitly manipulate the reverberation present to favor their preferred conclusion about ensemble size: compare Joshua Rifkin's (relatively dry) 1987 recording of BWV 106 to Masaaki Suzucki's 1996 version, which instead features longer (artificial) reverberation and slower tempos. Surely this is begging the question: certainly smaller ensembles benefits from a more intimate acoustic, while larger choirs can make better use of a long reverberant tail, but what acoustic conditions was Bach really working with? Room acoustic conditions profoundly affect composers and performers: for instance, Bach's role model Dietrich Buxtehude composed for the much larger church of St. Mary's in Lubeck, where his choir was forced to sing in unison to guide the congregation through the space's much longer reverberation time [17].

Bach was an organist and consultant with decades of experience in carefully listening to the acoustics of different churches. He spent the last 27 years of his life in Leipzig, composing and leading music for all the major liturgical and civi spaces of Leipzig. Bach had a strong preference for the Thomaskirche, possibly because he deemed it superior for choral music [18]. The combination of Bach's keen ear and his close association with the Thomaskirche make this church one of the most significant performance spaces in the history of Western music. However, the church has been altered many times, and its current interior is quite different from Bach's time [19].

In 1930 English acoustician Hope Bagenal, along with Bach scholar C.S. Terry, put forth the hypothesis that the relatively clear acoustics of the Thomaskirche were established by renovations during the Lutheran Reformation in the 16th century [20, 21]. Their primary argument was that the Lutherans' removal of screens (ensuring clear sight lines) and addition of seating galleries would have increased the musical clarity and reduced the church's reverberation time, creating a more clear acoustic channel for Bach's complex polyphonic music. Bagenal's 1930 article provided some rudimentary calculations (based on Sabine's formula) to argue for his conclusions. This study aims to put his hypothesis to the test by simulating the effect of the Reformation alterations on performers singing in the Bach-era church through a real-time auralization.

2.2 History of the Thomaskirche

Though Bagenal's narrative of the church's history was a bit simplistic, he did correctly name the 16th century alterations as the most relevant to the soundscape during Bach's tenure there. The earliest known construction on the site was a late-Gothic church dating from the mid-12th century, which experienced several different alterations from 1200-1500 [19]. Though the majority of the galleries were added after the Reformation, a smaller choir gallery existed at the west end of the nave as early as 1498 [22]. After Martin Luther preached in the church in 1539, early bouts of iconoclasm resulted in the removal of the rood screen from the chancel and the destruction of all but one of the church's fifteen altars. Due to the increased importance of the sermon in the Protestant service, the church required additional seating (note that the city did not state any acoustical motivation for this change, only the desire, common to many concert halls, to cram in as many listeners as possible). Thus in 1570-71 Hieronymous Lotter supervised the expansion of the west gallery (in blue on Fig. 1) for the Thomanerchor to occupy, and built the "Renaissance" galleries on the south and north of the nave, highlighted on Fig. 1 in brown [23].

These stone galleries exist today in part, but drawings such as that of Lilo Häring (Fig. 2) show that these were originally double-level galleries, similar to those still present in Leipzig's Nicholairkirche today [23]. The upper-level galleries were constructed entirely from wood while providing wooden pews full of congregants [24], which would have certainly increased the acoustic absorption compared to the pre-Reformation church with much lower seating capacity.

By Bach's time the upper galleries themselves stretched around to the sides of the west choir gallery, elevated "town piper galleries," (Fig. 3) [1] built in 1632, elevated above both side of the vocalists [23, 24]. The strings played from the south gallery, and the winds from the north gallery, while the continuo, trumpet, and drums (if any) were positioned around the organ's *Rückpositiv* [24]. Bach himself generally led from the south gallery, playing the violin [16].

3. METHOD

3.1 Acoustic Measurement and Simulation

The current project uses a combination of acoustic measurements in the current church and calibrated geometric acoustical (GA) modeling to simulate the acoustics of the church today and as it existed in 1723 (the beginning of Bach's tenure there) as well as the 1539 church with its lower seating capacity, rood screen, and intact altars. The church's room impulse response was measured at the source and receiver positions shown in Fig. 1 using a custom-built dodecahedral loudspeaker, a CoreSound TetraMic, and an exponential swept sinusoid [25]. Full details of these measurements will be addressed more fully in a future publication, but the empty church's mid-frequency T30 value was about 3.7s, similar to that measured before the church's renovations in the 1960s [26].

A GA model of the empty present-day church was constructed in CATT-Acoustic v.9.1 [27, 28]. The acoustic measurements were used to calibrate the material scattering and absorption coefficients within known surface values [29, 30]. Using this calibrated present-day model, two additional models were created: one for the pre-Reformation era church (c. 1539), and one for the post-Reformation church (c. 1723), corresponding to the state of the space when Bach assumed the position of cantor

[1] In private conversation Christoph Wolff has confided to the authors that he is revising this image to ensure that the lowest part of the piper galleries were even with the top of the lower gallery, unlike in Figure 3.

doi:10.25836/sasp.2019.39

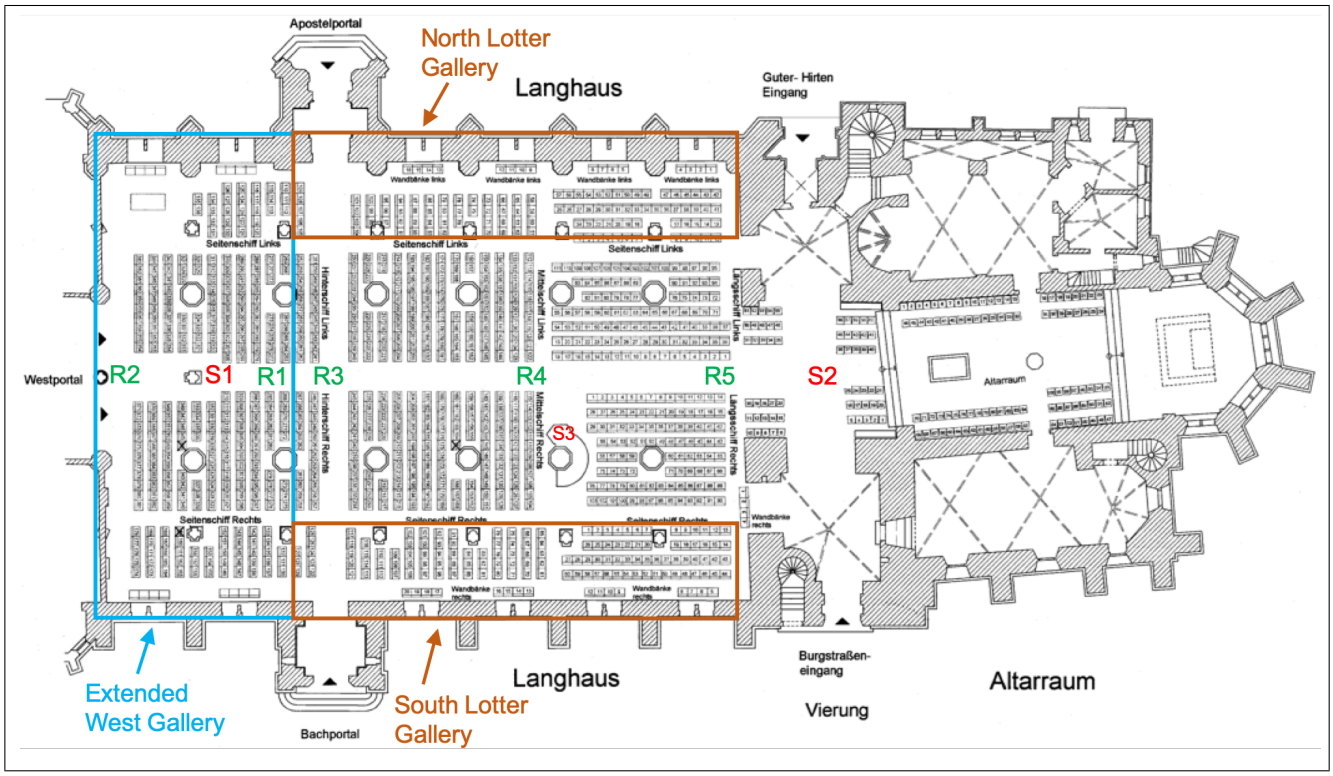


Figure 1. Plan of Thomaskirche, with positions of Lotter galleries and source/receiver positions during impulse response measurements.

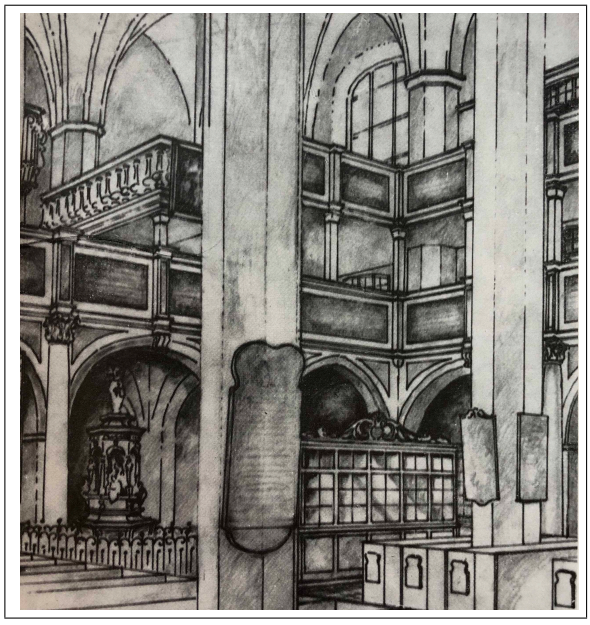


Figure 2. Reconstruction of the second level of galleries that existed in Bach's time [23].



Figure 3. Wolff's reconstruction of the town piper galleries around the west choir gallery [24]. The strings (and Bach) would have been in the south gallery (on the left), the woodwinds would have played from the north gallery opposite, with the choir one level down on the larger west gallery.

doi:10.25836/sasp.2019.39

there.

To create this calibrated GA model of the church in its present form, we started by modeling the church's outer geometry directly in CATT. This geometry is based on measurements taken from drawings of the church which we then verified on site with laser-based distance measurements. With the outer geometry of the church constructed, we added the surrounding galleries and pews, mapping seating areas as audience planes. Lastly, we assigned preliminary absorption and scattering coefficients to all surfaces. While some surfaces were common and had known scattering and absorption coefficients, others were unknown and required estimation and calibration. For the absorption coefficients of unknown materials, we began by using absorption data of similar materials with known absorption. Scattering coefficients were calculated using CATT's estimate function [30], given by

$$\text{scatt}(f)\Big|_{\geq 0.10}^{\leq 0.99} = 0.5\sqrt{\frac{\text{d}}{\lambda}}, \tag{1}$$

where λ is wavelength and d is the characteristic depth of the surface in meters.

3.2 Cantata Selection and Forces

The cantata *Herz und Mund und Tat und Leben*, BWV 147 ("Heart and Mouth, and Deed, and Living") was selected as the initial test piece for the process. Chosen for its variety of instrumental and solo vocal forces, the work features a highly contrapuntal *da capo* opening chorus, plus iterations of the well-known chorale setting – commonly known as "Jesu Joy of Man's Desiring" – as the sixth movement and with additional verses set as the final tenth movement of the work. It is a fairly rare work within the Bach oeuvre in its inclusion of solo arias for all four voices types (soprano, alto, tenor, and bass). Additionally, the solo movements and recitatives feature an astounding array of obbligati instruments including oboe d'amore, violin, organ, trumpet, and two oboe da caccia.

Originally composed for Weimar in 1716 (BWV 147a), the work was adapted and expanded for use during the Feast of the Visitation of the Blessed Virgin Mary in 1723 as part of Bach's first annual cantata cycle for Leipzig. The majority of movements are in simple or compound triple meter with the opening chorus notably consisting of 66 measures; the structures allude to the Holy Trinity and illustrate Bach's interest in numerology and his interest to connecting musical form to theology.

The recording was made with early-music specialists all performing on period instruments and bows. For the recording process, the choruses and chorales were recorded both following the "traditional choir" approach using four singers to a part as well as the "OVPP" (one vocalist per part) understanding alongside varied string forces with a larger tutti ensemble of three violin I, three violin II, two violas, two violoncelli, and one violone (eleven strings) and a smaller string configuration of 2-2-1-1-1 (seven strings) respectively. The opening chorus and chorales movements do not lend themselves easily to concertist and ripienist contrasts and additionally only a single set of solo vocal parts survive, eliminating the necessity to consider this kind of deployment. [2] Solo movements, when involving strings (Nos. 2 and 9) were recorded using the smaller string configuration.

3.3 Cantata Recording

Using this historical model, we have produced virtual sources for the primary musician groups present in Bach's time, including strings, woodwinds, continuo, trumpet, organ, and the choir. These sources have been placed in the GA model of the 1723 Thomaskirche's double galleries (which no longer exist) based on historical evidence of their locations during Bach's cantata performances. Each of these 6 sources was also used as a receiver, and the GA was used to generate binaural impulse responses for the 6x6 matrix of source-receiver combinations. In addition, three congregational source positions were simulated at different distances from the musicians along the nave of the church.

Since isolation was crucial for the accurate generation of the simulated acoustic environment, the microphones used were mainly directional condensers, and placed for the most optimal compromise between isolation, proximity effect and comb-filtering.

Miking the instruments individually also added to the overall complexity of the setup and utilization of the available physical space. For this reason the main control room in Kreeger Studios was used as an iso booth for the singers, with a video feed from the live room. The session was then recorded from a smaller API1608-based control room, with the addition of eight Grace Design preamps, for a total track count of 34 channels. The other instruments were placed in the live room with the conductor.

The software used was Pro Tools Ultimate, with HDX/HD I/O converters set to a sample rate of 48kHz and 24-bit. Higher sample rates were not used in order to allow for sends to multiple instances of convolution reverb plugins with minimum latency and CPU-processing. The simulated binaural impulse responses for each section-to-section path were then sent out from the reverb channels in Pro Tools to the musicians' headphones.

The recording sessions took place in August 2019 with the main ensemble in a dry recording studio, with additional absorbing surfaces placed between sections, reducing but not eliminating bleed entirely. Using this setup, different musicological variables (i.e. small versus larger instrumental and vocal forces) and acoustical variables (i.e. the BRIR from 1723 or 1539) could be tested during multiple recordings of the same chorale. This method allows the final auralization to incorporate the effect of the church's acoustics on the performance as well as the response at the audience's position.

Further discussion is needed for an optimal method for

[2] See A. Parrott, The Essential Bach Choir [14], for a summary of the use of *concertists* and *ripienists* (ch. 4) and for a discussion of the various views on instrumental forces (ch. 9).

doi:10.25836/sasp.2019.39

recording large ensembles dependent on the latency and bleed present in different setups, as well as the infrastructure needed to simultaneously record multiple performers in multiple rooms. In addition, as the project also involved the goal of creating an artistically sound final product, some tradeoffs had to be made between purely experimental virtual setups for shorter excerpts and finding stable positions where instrumentalists could hear each other clearly enough to play the entire cantata in the constrained studio time allowed.

4. DISCUSSION

Though it is well established that performers' tempo and articulation are shaped by the acoustics of their performance space [10], the question of how to recreate historical acoustical factors for ensemble performance has not yet been fully answered. Ideally we would like to have anechoic recordings of each instrument [31], but attempts to do this require a 'scratch track' that musicians play along with, which pre-determines the recording's tempo and reduces the feeling of realistic live performance for the musicians. The Virtual Haydn project, in contrast, provided virtual rooms according to historical simulations, but since only one (keyboard) player was needed, the recording could be conducted in a dry environment with no bleed [3]. In this case, though we did not have access to separate anechoic recording environments for each instrument, the use of a full recording studio allowed sufficient isolation for louder instruments, minimizing bleed while providing a very low noise floor. Close-miking each instrument allows each to be auralized separately, a technique which has provided good results even for fully reverberant environments [32].

While space constraints sometimes led to tradeoffs in player comfort during the recording process, this setup generally allowed the ensemble to experience the virtual space at two distinct moments in its history and adjust their performance accordingly. Although it would have been possible to simulate every individual instruments' binaural impulse response as heard at each musician's position, the amount of realtime convolutions necessary for this raised the risk of increased latency, as well as the possibility of crashing the ProTools setup; as a result the "section by section" approach was deemed good enough to render the macro-changes in the church's acoustics while still accounting for the spatial distances between the sections. Closed-back headphones (though not the preference of most working classical musicians) allowed the performers to hear the virtual space rather than the recording studio, while also minimizing bleed.

5. CONCLUSIONS AND FUTURE WORK

While the use of virtual models cannot replace documentary evidence in support of a musicological conclusion, so-called Computational Musicology (and the Digital Humanities more generally) can provide new kinds of evidence which must be interpreted within a larger musicological framework. The twisting historical path of Bach's influences, from religion to architecture to music can only be investigated through an acoustical lens, which may bring to light new aspects of the composer's practical limitations and aesthetic decisions. As VR technology and room modeling become more widespread, it is hoped that such techniques will also inform the standard practice in early music performance and recording.

The recordings from these sessions will have final reverb filters applied directly in CATT for a series of source-receiver auralizations at different listening points in the virtual Thomaskirche. Working with web and UX designers, the next step in the project will involve a user interface allowing listeners to experience the cantata from a series of listening points on the floor or in the galleries.

Future investigations will include other spaces in Bach's performance career and various genres of concerted pieces. Some acoustical aspects, such as diffraction effects, which are not adequately rendered by the GA model, can be modeled using a hybrid system employing Finite Difference modeling at low frequencies and ray tracing at high frequencies. In addition, future recordings may be conducted in a dry sound stage with more options for reconfiguring the ensemble to allow better isolation of each instrument while still rendering the real-time convolutions in such a way as to create a realistic feeling that the musicians are playing together in the present, but also in the past.

6. ACKNOWLEDGEMENTS

This research was funded by United States National Endowment for the Humanities' Digital Humanities grant award number A19-0069-001, "Hearing Bach's Music as Bach Heard It." The authors would like to thank Jackson Anthony for assistance in translation of many of the German architectural records and Davide Bonsi for assistance during the acoustic measurement of the Thomaskirche. Thanks also to Markus Rathey, Christoph Wolff, and Markus Zepf for their expertise on the architectural history of the Thomaskirche. Final thanks to Daniel Melamed for conversations and feedback on Bach's ensembles.

7. REFERENCES

[1] M. S. Cuthbert, "Notre Dame can be rebuilt, but its unique sound may be gone forever," *Los Angeles Times*, 2019.

[2] B. N. J. Postma and B. F. G. Katz, "Acoustics of Notre-Dame cathedral de Paris," in *22nd International Congress on Acoustics*, (Buenos Aires, Argentina), 2016.

[3] T. Beghin, M. de Francisco, and W. Woszczyk, "The Virtual Haydn," 2009.

[4] S. Weinzierl, H. Rosenheinrich, J. Blickensdorff, M. Horn, and A. Lindau, "Die Akustik der Konzertsäle im Leipziger Gewandhaus. Geschichte, Rekonstruktion und Auralisation," *Daga 2010*, no. 1781, pp. 1045–1046, 2010.

[5] B. B. Boren and M. Longair, "Acoustic simulation of the church of San Francesco della Vigna," in *Proceedings of Meetings on Acoustics: 164th Meeting of the Acoustical Society of America*, vol. 18, 2012.

[6] B. B. Boren, M. Longair, and R. Orlowski, "Acoustic Simulation of Renaissance Venetian Churches," *Acoustics in Practice*, vol. 1, no. 2, pp. 17–28, 2013.

[7] M. Sender, A. Planells, R. Perelló, J. Segura, and A. Giménez, "Virtual acoustic reconstruction of a lost church: application to an Order of Saint Jerome monastery in Alzira, Spain," *Journal of Building Performance Simulation*, vol. 11, no. 3, pp. 369–390, 2018.

[8] B. N. J. Postma, S. Dubouilh, and B. F. G. Katz, "An archeoacoustic study of the history of the Palais du Trocadero (18781937)," *The Journal of the Acoustical Society of America*, vol. 145, no. 4, pp. 2810–2821, 2019.

[9] B. B. Boren, "Computational acoustic musicology," *Digital Scholarship in the Humanities*, vol. 34, no. 4, 2019.

[10] T. Fischinger, K. Frieler, and J. Louhivuori, "Influence of Virtual Room Acoustics on Choir Singing," *Psychomusicology: Music, Mind, and Brain*, vol. 25, no. 3, pp. 208–218, 2015.

[11] J. Butt, *Playing with History: The Historical Approach to Musical Performance*. Cambridge, UK: Cambridge University Press, 2002.

[12] J. Rifkin, "Bach's Chorus," *The Musical Times*, vol. 123, no. 1677, pp. 747–751, 753–754, 1982.

[13] T. Koopman and L. Carolan, "Bach's Choir, an Ongoing Story," *Early Music*, vol. 26, no. 1, pp. 109–121, 1998.

[14] A. Parrott, *The Essential Bach Choir*. Woodbridge, UK: Boydell Press, 2000.

[15] R. A. Leaver, "Performing Bach: One or Many?," *The Choral Scholar*, vol. 1, no. 1, pp. 6–15, 2009.

[16] B. Jerold, "Performance conditions, standards and Bach's chorus," *The Musical Times*, vol. 158, no. 1941, pp. 55–70, 2017.

[17] K. J. Snyder, "Tradition with Variations: Chorale Setting per omnes versus by Buxtehude and Bach," in *Music and Theology: Essays in Honor of Robin A. Leaver* (D. Zager, ed.), ch. 3, pp. 31–50, Plymouth, UK: Scarecrow Press, 2007.

[18] H. T. David, A. Mendel, and C. Wolff, eds., *The New Bach Reader*. New York and London: W. W. Norton and Company, 1998.

[19] C. Wolff, "Building and Construction History," in *St. Thomas Church in Leipzig* (B. Taddiken, ed.), pp. 5–14, Leipzig, Germany: Evangelische Verlagsanstalt, 2017.

[20] H. Bagenal, "Bach's Music and Church Acoustics," *Music and Letters*, vol. 11, no. 2, pp. 146–155, 1930.

[21] C. S. Terry, "Bach's Music and Church Acoustics," *The Consort*, vol. 2, pp. 2–3, 1931.

[22] M. Petzoldt, *St. Thomas Zu Leipzig*. Leipzig, Germany: Evangelische Verlagsanstalt, 2000.

[23] H. Stiehl, "Das Innere der Thomaskirche zur Amtszeit Johann Sebastian Bach," *Beiträge zur Bachforschung*, vol. 3, 1984.

[24] C. Wolff, *Johann Sebastian Bach: the learned musician*. New York, NY: W. W. Norton and Company, 2000.

[25] A. Farina, "Simultaneous Measurement of Impulse Response and Distortion with a Swept-Sine Technique," in *Proceedings of the 108th Audio Engineering Society Convention*, (Paris), 2000.

[26] L. Keibs and W. Kuhl, "Zur Akustik der Thomaskirche in Leipzig," *Acustica*, vol. 9, pp. 365–370, 1959.

[27] B.-I. Dalenbäck, "Room Acoustic Prediction Based on a Unified Treatment of Diffuse and Specular Reflection," *Journal of the Acoustical Society of America*, vol. 100, no. 2, pp. 899–909, 1996.

[28] B.-I. Dalenbäck, *CATT-Acoustic v9*. Gothenburg, Sweden: CATT, 2011.

[29] B. B. Boren and M. Longair, "A Method for Acoustic Modeling of Past Soundscapes," in *Proceedings of the Acoustics of Ancient Theatres Conference*, (Patras, Greece), 2011.

[30] B. N. J. Postma and B. F. G. Katz, "Creation and calibration method of acoustical models for historic virtual reality auralizations," *Virtual Reality*, vol. 19, no. 3-4, pp. 161–180, 2015.

[31] J. Pätynen, V. Pulkki, and T. Lokki, "Anechoic recording system for symphony orchestra," *Acta Acustica united with Acustica*, vol. 94, no. 6, pp. 856–865, 2008.

[32] B. N. J. Postma, D. Poirier-Quinot, J. Meyer, and B. F. Katz, "Virtual Reality Performance Auralization in a Calibrated Model of Notre-Dame Cathedral," in *EuroRegio2016*, (Porto, Portugal), 2016.

 doi:10.25836/sasp.2019.39

AN EVALUATION OF PRE-PROCESSING TECHNIQUES FOR VIRTUAL LOUDSPEAKER BINAURAL AMBISONIC RENDERING

Thomas McKenzie, Damian T. Murphy, Gavin Kearney
AudioLab, Communication Technologies Research Group,
Department of Electronic Engineering, University of York,
York, YO10 5DD, UK
thomas.mckenzie@york.ac.uk

ABSTRACT

Binaural Ambisonic rendering can be achieved using virtual loudspeakers through head-related impulse response (HRIR) convolution of the Ambisonic loudspeaker feeds. It is widely used in immersive applications such as virtual reality due to its sound field rotation capabilities and low channel count. Binaural Ambisonic reproduction is inaccurate at high frequencies, causing reduced localisation and timbral accuracy, but can be improved through offline pre-processing of the virtual loudspeaker HRIRs. This paper details a numerical and perceptual evaluation of several state of the art pre-processing technique combinations.

1. INTRODUCTION

Binaural Ambisonic rendering is well suited to virtual reality applications due to its sound field rotation capabilities. Ambisonic reproduction can theoretically replicate the original sound field exactly in the region of the head for frequencies up to what is commonly referred to as the 'spatial aliasing frequency', f_{alias}, but at frequencies above f_{alias}, reproduction can become inaccurate due to the limited spatial accuracy of reproducing a physical sound field with a finite number of transducers, which in practice causes localisation blur [8], reduced lateralisation [1] and comb filtering spectral artefacts [2].

The standard approach to improving Ambisonic reproduction is to increase the order of Ambisonics, which allows for exact sound field reproduction up to a higher f_{alias} [3, 4], though at the expense of more channels for storage, more microphone capsules for recording, and more convolutions in binaural reproduction. It is therefore highly desirable to explore alternative methods of improving low-order Ambisonic rendering. One common practice is to employ a dual-band decoder with basic Ambisonic decoding at low frequencies and Max r_E channel weighting above f_{alias} [1, 5], which improves spectral, localisation and lateralisation reproduction at high frequencies.

© Thomas McKenzie, Damian T. Murphy, Gavin Kearney. Licensed under a Creative Commons Attribution 4.0 International License (CC BY 4.0). **Attribution:** Thomas McKenzie, Damian T. Murphy, Gavin Kearney. "An evaluation of pre-processing techniques for virtual loudspeaker binaural Ambisonic rendering", 1st EAA Spatial Audio Signal Processing Symposium, Paris, France, 2019.

This paper presents the method and results of numerical and perceptual evaluations of state-of-the-art pre-processing technique combinations for virtual loudspeaker binaural Ambisonic rendering. These pre-processing techniques are applied to the head-related impulse responses (HRIRs) used in the virtual loudspeaker binaural rendering stage in offline processes, such that the resulting binaural decoders are of the same size and require the same number of real-time convolutions. This paper investigates 1st, 2nd and 3rd order Ambisonics, with loudspeaker configurations comprising 6, 14 and 26 loudspeakers respectively, arranged in Lebedev grids [9].

All HRIRs used in this study were generic measurements from a diffuse-field equalised version of the Bernschütz Neumann KU 100 database [6]. All computation was carried out offline in MATLAB version 9.3.0 - R2017b and Ambisonic encoding and decoding was achieved using the Politis Ambisonic library [7], which uses three-dimensional full normalisation (N3D) and Ambisonic channel number (ACN) ordering. All audio used was of 24-bit depth and 48 kHz sample rate. Ambisonics was rendered using mode-matching pseudo-inverse decoding and, unless otherwise stated, dual-band decoding was utilised with basic channel weightings at frequencies below f_{alias} and Max r_E weightings above [1, 5]. In this study, f_{alias} was approximated according to [8] with a speed of sound of 343 m/s and radius of the listening area as 9 cm (the approximate radius of the Neumann KU 100 dummy head) as 670 Hz, 1270 Hz and 1870 Hz, for 1st, 2nd and 3rd order Ambisonics, respectively.

2. PRE-PROCESSING TECHNIQUES

The three pre-processing techniques tested in this paper are time alignment (TA), Ambisonic interaural level difference optimisation (AIO) and diffuse-field equalisation (DFE). TA is the complete removal of interaural time differences (ITDs) of the HRIRs at high frequencies [10,11], which reduces the comb filtering caused by the off-centre position of the ears in the virtual loudspeaker array. TA has previously only been implemented for dense sets of HRIRs and is here applied to sparse virtual loudspeaker sets. When using TA, basic channel weighting is usually used for the whole frequency spectrum. The crossover frequency above which TA is implemented in this study is 2.5kHz; chosen

doi:10.25836/sasp.2019.09

according to the listening test results in [12]. AIO is an iterative pre-processing technique that brings the Ambisonic rendering of interaural level differences (ILDs) [13] closer to that of HRIRs. It is achieved by augmenting the high frequency magnitude of virtual loudspeaker HRIRs such that when rendering Ambisonics with the augmented HRIRs, the ILD is reproduced with greater accuracy. DFE is the removal of direction-independent spectral artefacts in the Ambisonic rendered diffuse-field [14], which improves the overall spectral reproduction of binaural Ambisonic rendering, as well as the timbral consistency between different virtual loudspeaker configurations [15]. It works by creating Ambisonic renders at a large amount of directions on the sphere and obtaining an average of all the frequency responses. This is then inverted and the equalisation filter is applied to the virtual loudspeaker HRIRs.

In this paper, different pre-processing techniques are combined. Theoretically, by running one after the other, the resulting binaural Ambisonic decoder will produce greater results than just one of the pre-processing techniques. The order of pre-processing techniques is as follows: TA is implemented first as it affects the rendering of ILD and the diffuse-field response. AIO also affects the diffuse-field response, so follows TA. DFE is implemented last, as it corrects any changes in average frequency response and the other pre-processing techniques can affect the diffuse-field response. The five binaural Ambisonic decoders under test in this paper (along with their abbreviations), with order of implementation, are as follows:

- NPP: Standard Ambisonic (dual band)
- PP 1: AIO & DFE (dual band)
- PP 2: TA & DFE (basic)
- PP 3: TA & AIO & DFE (basic)
- PP 4: TA & AIO & DFE (dual band)

AIO produces the greatest benefits for dual-band decoding, but TA is recommended for basic weighted decoding. Therefore, in PP 3 and PP 4 with the combination of all three pre-processing techniques, both basic weighted and dual-band instances are included to ascertain the differences and determine which (if any) is superior.

3. NUMERICAL EVALUATION

To assess the effect of different pre-processing combinations on the spectral accuracy of binaural Ambisonic rendering, a perceptually motivated spectral difference (PSD) fast Fourier transfer (FFT) based model was used [16]. It weights input signals using ISO 226 equal loudness contours and a sone scale to account for the loudness-varying sensitivity of human hearing as well as equivalent rectangular bandwidth weightings to address the linear frequency sample spacing of FFTs.

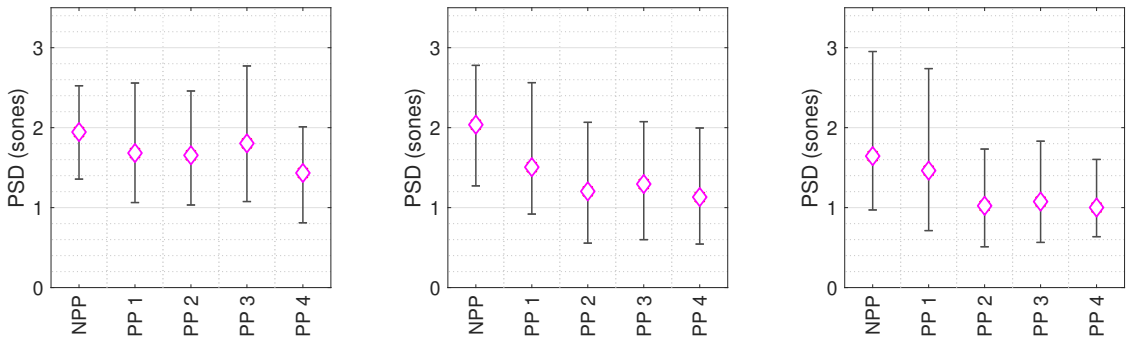
PSD between binaural Ambisonic renders and HRIRs for 16,020 locations over the sphere (distributed using a 2° Gaussian grid) with different combinations of pre-processing techniques for 1st, 2nd and 3rd order Ambisonics was calculated. Figure 1 shows the solid angle weighted average PSD value for each pre-processing combination for all tested orders of Ambisonics, with whiskers to denote the maximum and minimum PSD values. As expected, higher orders of Ambisonics produce improved spectral reproduction. In all tested orders of Ambisonics, every pre-processing combination improves the overall spectral accuracy over standard dual-band decoding, but PP 4 (the dual-band combination of TA, AIO and DFE) produces the greatest improvements with the lowest solid-angle weighted PSD and the lowest value of maximum PSD for all 3 tested orders of Ambisonics. To illustrate how PSD changes over the sphere, the PSD for each combination of 1st order pre-processing techniques for all locations on the sphere is presented in Figure 2 (to conserve space, 2nd and 3rd order plots are omitted).

To assess the effect of different pre-processing combinations on the accuracy of ILD reproduction in binaural Ambisonic rendering, the ILD between binaural Ambisonic renders and HRIRs for 16,020 directions over the sphere with different pre-processing combinations was estimated using the method in [17]. The solid angle weighted change in ILD (referred to here on in as ΔILD) between HRIRs and binaural Ambisonic renders with different pre-processing combinations for all directions on the sphere and all tested orders of Ambisonics is presented in Figure 3, with whiskers to denote the maximum ΔILD value. As expected, higher orders of Ambisonics produce improved ILD rendering, and in all tested orders of Ambisonics, every pre-processing combination improves the solid-angle weighted ΔILD over standard dual-band decoding. Interestingly however, different orders produce different results for which pre-processing combination offers the best ILD reproduction. For 1st and 3rd order, PP 4 (the dual-band combination of TA, AIO and DFE) produces the greatest improvements with the lowest solid-angle weighted ΔILD and the lowest value of maximum ΔILD, but for 2nd order, this is found at PP 3 (the basic channel weighted combination of TA, AIO and DFE). To illustrate how ΔILD changes depending on the location on the sphere, Figure 4 shows ΔILD for all locations on the sphere and each pre-processing combination for 1st order Ambisonics (to reduce the overall amount of figures, 2nd and 3rd order plots are omitted).

4. PERCEPTUAL EVALUATION

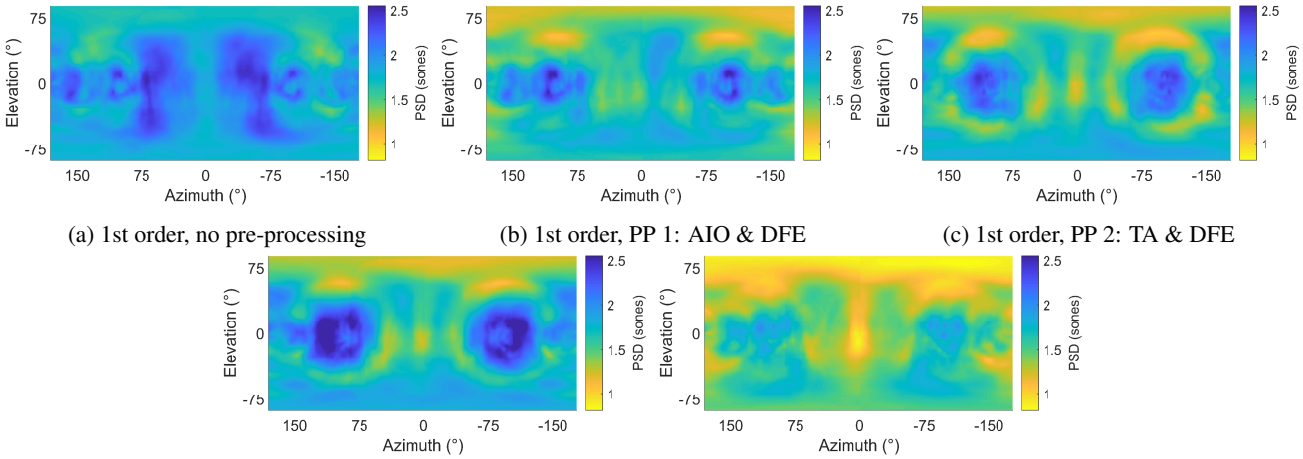
To assess the perceptual effect of different pre-processing combinations, listening tests were conducted using both simple and complex acoustic scenes. The tests followed the multiple stimulus with hidden reference and anchors (MUSHRA) paradigm, ITU-R BS.1534-3 [18]. Tests were conducted in a quiet listening room using an Apple Macbook Pro with a Fireface 400 audio interface, which has software controlled input and output levels. A single set of Sennheiser HD 650 circum-aural headphones were used, which were equalised using Kirkeby and Nelson's least-

doi:10.25836/sasp.2019.09



(a) 1st order (b) 2nd order (c) 3rd order

Figure 1: Solid angle weighted PSD values between 16,020 HRIRs and binaural Ambisonic renders with different pre-processing combinations, for 1st, 2nd and 3rd order Ambisonics. Whiskers denote maximum and minimum PSD values and NPP denotes no pre-processing.



(a) 1st order, no pre-processing (b) 1st order, PP 1: AIO & DFE (c) 1st order, PP 2: TA & DFE

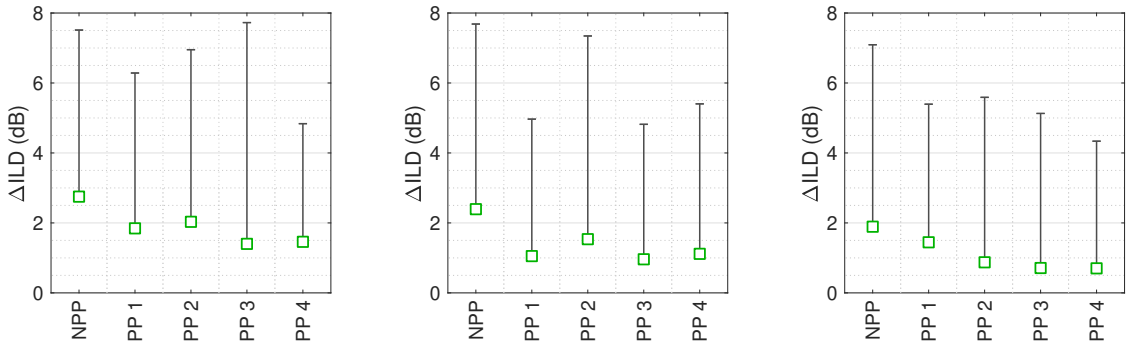
(d) 1st order, PP 3: TA & AIO & DFE (basic) (e) 1st order, PP 4: TA & AIO & DFE (dual band)

Figure 2: Perceptual spectral difference between 1st order binaural Ambisonic renders and HRIRs over the sphere with different pre-processing combinations.

mean-square regularization method [19] from the RMS average of 11 impulse response measurements collected using Farina's swept sine technique [20] and a Neumann KU 100. The range of inversion was 5 Hz–4 kHz, and in / out-band regularization of 25 dB and -2 dB respectively was employed to avoid sharp peaks in the inverse filters. 20 experienced listeners took part, aged between 22 and 41, with no reported hearing impairments.

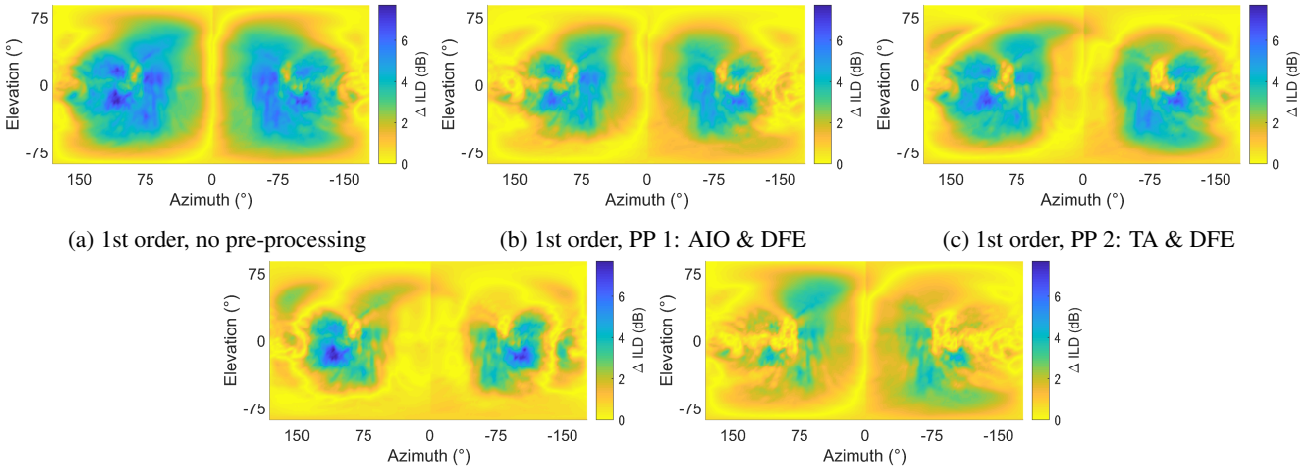
4.1 Test Methodology

Tests were conducted using static binaural rendering with no head-tracking implemented. Listeners compared binaural Ambisonic renders created using the pre-processing combinations as throughout this paper. All scenarios were repeated once. Three types of stimuli were used in the listening test. The first was a pseudo-moving pink noise sound, generated using 45 bursts of pink noise played consecutively and lasting 0.05 seconds long each, panned between $(\theta, \phi) = (44°, 0°)$ and $(\theta, \phi) = (132°, 0°)$ in 2° increments. Each burst was windowed using a 50 sample hanning window, resulting in a full pseudo-moving sound lasting 2.25 seconds. The reference was made from direct HRIR convolutions, and a monophonic version of the HRIR reference low-passed at 3.5 kHz was used as the low anchor. The second was a synthesised complex scene which comprised of 8 monophonic percussion tracks panned to 8 of the centre vertices of the faces of a dodecahedron. The reference was created by summing direct HRIR convolutions of the 8 original monophonic tracks, and again a monophonic version of the reference low-passed at 3.5 kHz was used as the low anchor. The third stimuli type was a 5 second excerpt of a fourth-order Ambisonic recordings of a beach soundscape made using an mh acoustics em32 Eigenmike [21]. The recording was converted from Schmidt semi-normalised (SN3D) to N3D normalisation using the method in [13]. As the Eigenmike recording test could not use an HRIR render as a reference, listeners were in this case asked to rate the stimuli in terms of plausibility, which was defined as 'a simulation in agreement with the listener's expectation towards a corresponding real event' [22]. An anchor was included as a monophonic version of the Ambisonic render with no



(a) 1st order (b) 2nd order (c) 3rd order

Figure 3: Solid angle weighted ΔILD between 16,020 HRIRs and binaural Ambisonic renders with different pre-processing combinations, for 1st, 2nd and 3rd order Ambisonics. Whiskers denote maximum ΔILD values and NPP denotes no pre-processing.



(a) 1st order, no pre-processing (b) 1st order, PP 1: AIO & DFE (c) 1st order, PP 2: TA & DFE

(d) 1st order, PP 3: TA & AIO & DFE (basic) (e) 1st order, PP 4: TA & AIO & DFE (dual band)

Figure 4: ΔILD between HRIRs and 1st order binaural Ambisonic renders over the sphere with different pre-processing combinations.

pre-processing.

4.2 Results

Listening test data was checked for normality using the one-sample Kolmogorov-Smirnov test, which showed all data as non-normal. Therefore, results were analysed using non-parametric statistics. Figures 5, 6 and 7 present the median scores for orders 1, 2 and 3 with non-parametric 95% confidence intervals [23] for the moving noise, percussion and beach stimuli types, respectively.

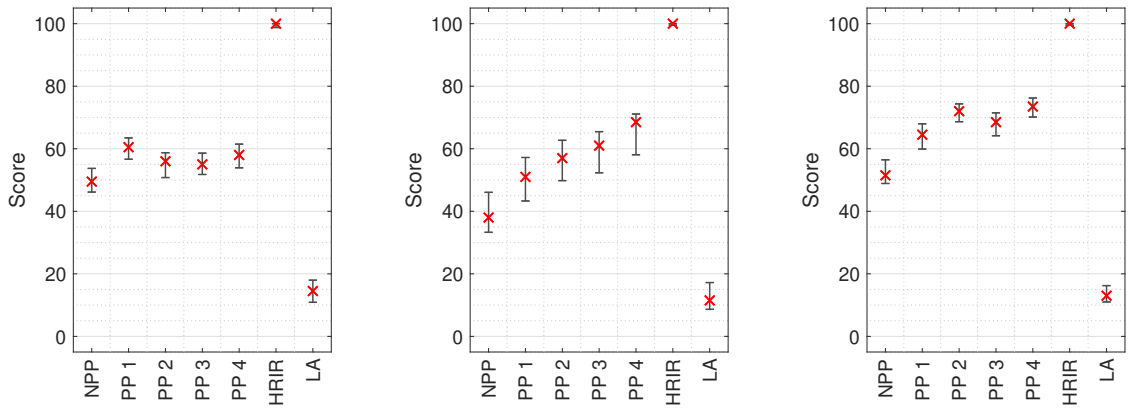
In all scenarios, NPP was rated as the worst Ambisonic condition. To assess the statistical significance of the differences between pre-processing combinations, Friedman's ANOVA tests were conducted on all test stimuli and orders. For the moving noise stimuli type, statistical significance was only found at 3rd order ($\chi^2(4) = 3.4, p = 0.5$; $\chi^2(4) = 6.3, p = 0.18$; $\chi^2(4) = 15.7, p < 0.01$ for 1st, 2nd and 3rd orders, respectively). For the percussion stimuli type however, this effect was significant for all tested orders ($\chi^2(4) = 19.7, p < 0.01$; $\chi^2(4) = 17.4, p < 0.01$; $\chi^2(4) = 34.2, p < 0.01$ for 1st, 2nd and 3rd orders, respectively). For the beach stimuli type, pre-processing combinations produced statistically significantly different results again only for 3rd order ($\chi^2(4) = 9.3, p = 0.05$; $\chi^2(4) = 7.3, p = 0.12$; $\chi^2(4) = 16.2, p < 0.01$ for 1st, 2nd and 3rd orders, respectively).

5. DISCUSSION AND CONCLUSIONS

This paper has presented an evaluation of pre-processing technique combinations for virtual loudspeaker binaural Ambisonic rendering. It is clear that pre-processing produces an improvement for all tested orders, something that has been shown both numerically and perceptually. However, results are not as simple as to offer a definitive optimal pre-processing combination and therefore warrant further discussion and testing.

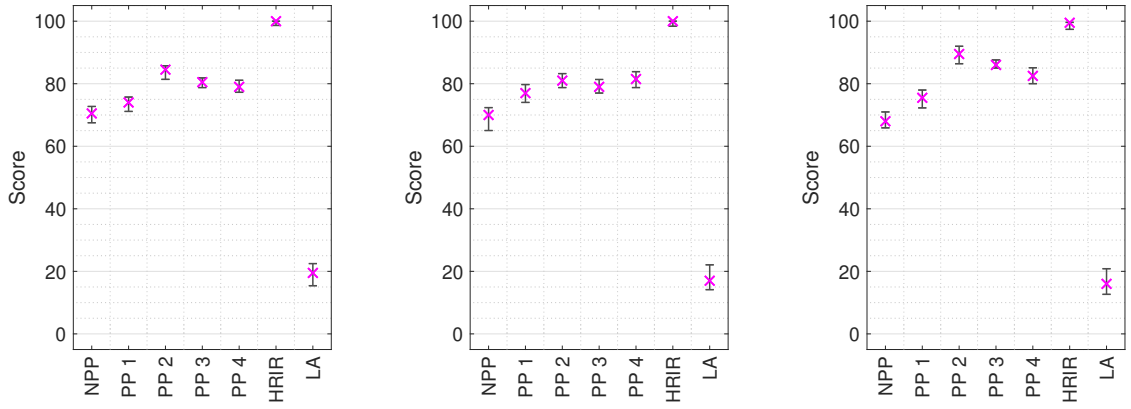
A discrepancy between the numerical and perceptual results for 1st order is notable where PP 4 clearly outperformed the other pre-processing combinations in spectral and ILD reproduction, but was not rated the highest for

doi:10.25836/sasp.2019.09



(a) 1st order (b) 2nd order (c) 3rd order

Figure 5: Median scores of the moving noise stimuli tests with non-parametric 95% confidence intervals. Scores indicate perceived similarity to the HRIR reference. LA and NPP denote low anchor and no pre-processing, respectively.



(a) 1st order (b) 2nd order (c) 3rd order

Figure 6: Median scores of the percussion stimuli tests with non-parametric 95% confidence intervals. Scores indicate perceived similarity to the HRIR reference. LA and NPP denote low anchor and no pre-processing, respectively.

any test stimuli type in the perceptual tests.

Perceptual results differed with test stimuli types. PP 2 performed better for the percussion stimuli type, whereas pre-processing combinations with AIO (PP 1, PP 3 and PP 4) performed better for the other two stimuli types. One possible explanation for this is that there was greater lateralisation present in the moving noise and soundscape stimuli.

The improvement gained from implementing TA has been shown to increase with order. This has been reflected in both the numerical and perceptual evaluation results, for all the pre-processing combinations that include TA (PP 2, PP 3 and PP 4), and is a likely factor for the statistical significance in variation of results for 3rd order with all stimuli types.

Future work will look at comparing the pre-processing technique combinations presented in this paper to other state of the art Ambisonic decoding solutions such as Magnitude Least Squares [12] and directional equalisation strategies [24].

6. REFERENCES

[1] J. Daniel, J.-B. Rault, and J.-D. Polack, "Ambisonics encoding of other audio formats for multiple listening conditions," in *105th Convention of the Audio Engineering Society*, 1998. Preprint 4795.

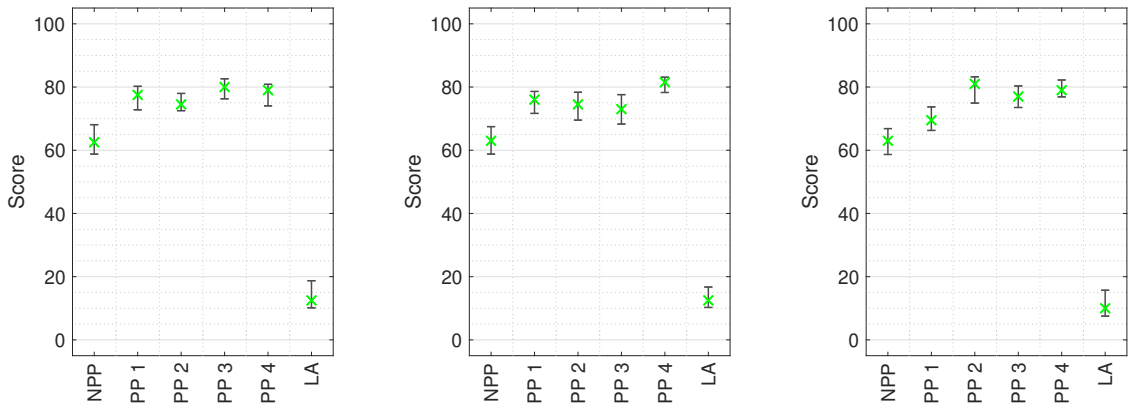
[2] J.-M. Jot, V. Larcher, and J.-M. Pernaux, "A comparative study of 3-D audio encoding and rendering techniques," in *AES 16th International Conference*, pp. 281–300, 1999.

[3] J. S. Bamford and J. Vanderkooy, "Ambisonic Sound for Us," in *99th Convention of the Audio Engineering Society*, 1995. Preprint 4138.

[4] D. G. Malham, "Higher order Ambisonic systems for the spatialisation of sound," in *Proc. ICMC 1999*, pp. 484–487, 1999.

[5] M. A. Gerzon and G. J. Barton, "Ambisonic decoders for HDTV," in *92nd Convention of the Audio Engineering Society*, 1992. Preprint 3345.

[6] B. Bernschütz, "A spherical far field HRIR/HRTF



(a) 1st order (b) 2nd order (c) 3rd order

Figure 7: Median scores of the beach stimuli tests with non-parametric 95% confidence intervals. Scores indicate perceived plausibility. LA and NPP denote low anchor and no pre-processing, respectively.

compilation of the Neumann KU 100," in *Fortschritte der Akustik – AIA-DAGA 2013*, pp. 592–595, 2013.

[7] A. Politis, *Microphone array processing for parametric spatial audio techniques*. PhD thesis, Aalto University, 2016.

[8] S. Bertet, J. Daniel, E. Parizet, and O. Warusfel, "Investigation on localisation accuracy for first and higher order Ambisonics reproduced sound sources," *Acta Acustica united with Acustica*, vol. 99, no. 4, pp. 642–657, 2013.

[9] V. I. Lebedev, "Quadratures on a sphere," *USSR Computational Mathematics and Mathematical Physics*, vol. 16, no. 2, pp. 10–24, 1976.

[10] M. J. Evans, J. A. S. Angus, and A. I. Tew, "Analyzing head-related transfer function measurements using surface spherical harmonics," *Journal of the Acoustical Society of America*, vol. 104, no. 4, pp. 2400–2411, 1998.

[11] M. Zaunschirm, C. Schörkhuber, and R. Höldrich, "Binaural rendering of Ambisonic signals by HRIR time alignment and a diffuseness constraint," *Journal of the Acoustical Society of America*, vol. 143, no. 6, pp. 3616–3627, 2018.

[12] C. Schörkhuber, M. Zaunschirm, and R. Höldrich, "Binaural Rendering of Ambisonic Signals via Magnitude Least Squares," in *DAGA 2018: 44. Deutsche Jahrestagung für Akustik*, pp. 339–342, 2018.

[13] T. McKenzie, D. Murphy, and G. Kearney, "Interaural level difference optimisation of binaural Ambisonic rendering," *Applied Sciences*, vol. 9, no. 6, 2019.

[14] T. McKenzie, D. Murphy, and G. Kearney, "Diffuse-field equalisation of binaural Ambisonic rendering," *Applied Sciences*, vol. 8, no. 10, 2018.

[15] T. Mckenzie, G. Kearney, and D. Murphy, "Diffuse-Field Equalisation of First Order Ambisonics," in *20th International Conference on Digital Audio Effects (DAFx-17)*, pp. 1–6, 2017.

[16] C. Armstrong, T. McKenzie, D. Murphy, and G. Kearney, "A perceptual spectral difference model for binaural signals," in *145th Convention of the Audio Engineering Society*, 2018. Convention E–Brief 457.

[17] K. Watanabe, K. Ozawa, Y. Iwaya, Y. Suzuki, and K. Aso, "Estimation of interaural level difference based on anthropometry and its effect on sound localization," *Journal of the Acoustical Society of America*, vol. 122, no. 5, pp. 2832–2841, 2007.

[18] ITU-R-BS.1534-3, "Method for the subjective assessment of intermediate quality level of audio systems BS Series Broadcasting service (sound)," vol. 3, 2015.

[19] O. Kirkeby and P. A. Nelson, "Digital filter design for inversion problems in sound reproduction," *Journal of the Audio Engineering Society*, vol. 47, no. 7/8, pp. 583–595, 1999.

[20] A. Farina, "Simultaneous measurement of impulse response and distortion with a swept-sine technique," in *108th Convention of the Audio Engineering Society*, 2000. Preprint 5093.

[21] M. Green and D. Murphy, "EigenScape: a database of spatial acoustic scene recordings," *Applied Sciences*, vol. 7, no. 12, 2017.

[22] A. Lindau and S. Weinzierl, "Assessing the plausibility of virtual acoustic environments," *Acta Acustica united with Acustica*, vol. 98, no. 5, pp. 804–810, 2012.

[23] R. Mcgill, J. W. Tukey, and W. A. Larsen, "Variations of box plots," *The American Statistician*, vol. 32, no. 1, pp. 12–16, 1978.

[24] T. McKenzie, D. Murphy, and G. Kearney, "Directional bias equalisation of first-order binaural Ambisonic rendering," in *AES Conference on Audio for Virtual and Augmented Reality*, 2018.

doi:10.25836/sasp.2019.09

COMPUTATIONAL MODELS FOR LISTENER-SPECIFIC PREDICTIONS OF SPATIAL AUDIO QUALITY

Piotr Majdak and Robert Baumgartner
Acoustics Research Institute
Austrian Academy of Sciences
piotr.majdak@oeaw.ac.at

ABSTRACT

Millions of people use headphones every day for listening to music, watching movies, or communicating with others. Headphones can be used to present binaural virtual sounds by filtering a sound with the head-related transfer functions (HRTFs). Here, we discuss aspects of spatial hearing that are particularly sensitive to listener-specific HRTFs, that is, sound localization along sagittal planes (i.e., vertical planes being orthogonal to the interaural axis) and near distances (sound externalization/internalization). We focus on recent findings aiming at predicting these two spatial audio qualities. We show that sagittal-plane localization is well understood and its models can reliably predict the localization performance. In contrast, more light needs to be shed onto the importance of the diversity of cues affecting sound externalization. To this end, we present results from a model-based meta-analysis of psychoacoustic studies. As potential cues we consider monaural and interaural spectral-shapes, spectral and temporal fluctuations of interaural level differences, interaural coherences, and broadband inconsistencies between interaural time and level differences in a highly comparable template-based modeling framework. Our investigations revealed that the monaural spectral-shapes and the strengths of time-intensity trading are potent cues to explain previous results under anechoic conditions.

1. INTRODUCTION

The acoustic basis for spatial hearing is formed by the acoustic filtering of sound sources by the listener's anatomy (see Fig. 1a). This filtering can be described by the listener-specific HRTFs [1], [2]. Acoustically measured or numerically calculated HRTFs of a listener can be used to filter a signal of a recorded or synthesized sound source in order to create virtual sources to be presented via headphones (see Fig. 1b).

When filtered with listener-specific HRTFs, the sounds presented via headphones can be indistinguishable from natural sounds [3]. In contrast, without listener-specific HRTFs, sounds presented via headphones are usually perceived inside the head instead of being localized at a naturally external position. Such an unnaturally perceived distance of sound sources has been coined "sound internalization", as opposed to the sound externalization perceived in natural listening situations [4]. Sound externalization is just a single dimension of the 3D sound localization. The other dimensions consider lateral halve planes (i.e., frontal or rear parts of the horizontal planes) and sagittal planes (i.e., vertical planes being orthogonal to the interaural axis), both forming the so-called interaural polar coordinate system [5], in which the direction of a sound source is described by the lateral and polar angles (see Fig. 2).

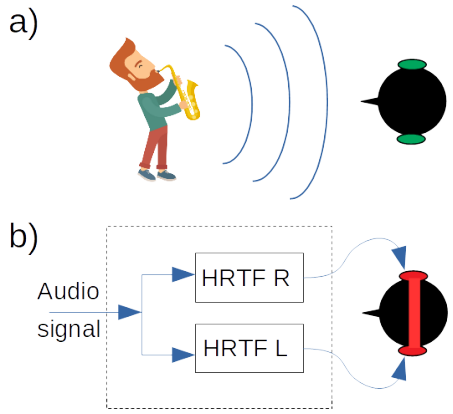


Figure 1. a) natural listening condition with listener's own ears (green). **b)** binaural sound reproduction by means of a binaural pair of HRTFs and headphones (red).

Note that besides 3D sound localization, spatial hearing also involves other perceptual attributes like apparent source width [6], listener envelopment [7], and the ability to segregate sounds.

In this contribution, we will focus on the dimensions of sound localization that are particularly sensitive to listener-specific HRTFs, that is, sagittal-plane sound localization and sound externalization.

2. SAGITTAL-PLANE SOUND LOCALIZATION

The sagittal-plane localization process is well understood. The auditory system processes monaural spectral features focusing on those that depend on the direction of a sound. Modern models aiming at predicting sound localization in sagittal planes consider the listener's HRTFs [8], [9]. We have recently developed a model which mimics the first relevant nucleus of the auditory system

© Piotr Majdak and Robert Baumgartner. Licensed under a Creative Commons Attribution 4.0 International License (CC BY 4.0). **Attribution:** Majdak, P., and Baumgartner, R. "Computational models for listener-specific predictions of spatial audio quality", 1st EAA Spatial Audio Signal Processing Symposium, Paris, France, 2019.

doi:10.25836/sasp.2019.10

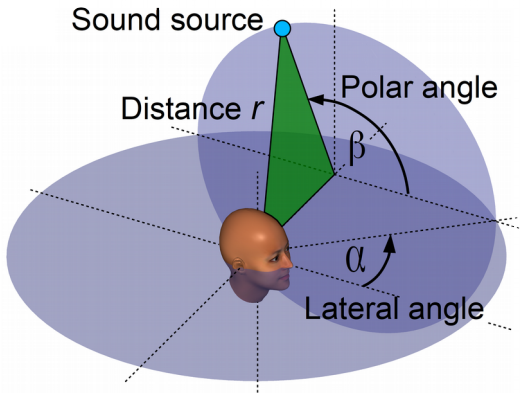


Figure 2. Sound-source localization within the interaural-polar coordinate system consisting of the lateral dimension (α, left/right), the sagittal dimension (β, up/down and front/back), and the distance (r).

known to process monaural spectral features, the dorsal cochlear nucleus (DCN). Neuronal networks within the DCN have been shown to be sensitive to rising spectral edges of the incoming sound [10]. Based on a simplification of this functionality, we have developed a model of sound localization in the sagittal planes [9], [11]. The incoming signal, after peripheral pre-processing, is mapped to positive spectral gradients, which are then compared with the templates representing the learned HRTFs (see Fig. 3). The model outputs the probability of responding to a polar angle, given an incoming sound and has been evaluated in various listening conditions.

As an example, results from listening with others' HRTFs are shown in Fig. 4. The actual results are replotted from [12] who tested 11 normal-hearing listeners localizing 250-ms long Gaussian white noise bursts either with their own listener-specific HRTFs or with HRTFs from other listeners. The model predictions were calculated for 23 listeners by calibrating their individual sensitivity parameter to their individual localization responses when listening to 500-ms long Gaussian white noise bursts. Then, without changing the model parameters, predictions for localizing with others' HRTFs were calculated for random target directions and summarized to quadrant errors and local polar root-mean-square (RMS) errors. The statistics of the predicted localization performance as represented by means, medians, and percentiles

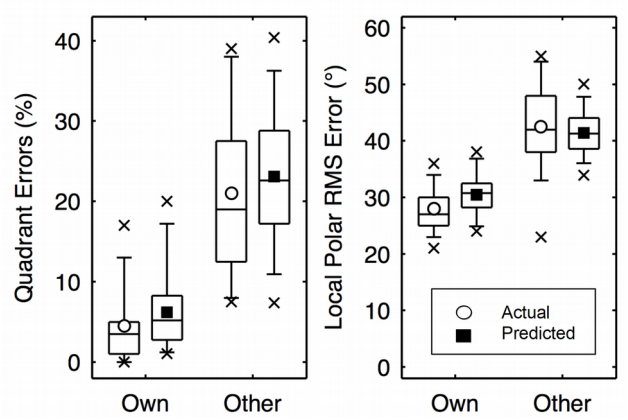


Figure 4. Modeling sagittal-plane sound localization for listening with own HRTFs and those from others. **Open symbols:** statistics of listener responses from the actual sound localization experiment (replotted from [12]). **Filled squares:** statistics of listener responses as predicted by the model. **Quadrant errors:** rate of hemifield confusions (front/back, top/bottom). **Local polar RMS error:** Root-mean-squared error between the listener response within the correct hemifield. **Statistics:** Circles and squares denote averages, horizontal lines represent 25th, 50th, and 75th percentiles, the whiskers represent 5th and 95th percentiles, and crosses represent minima and maxima.

appear to be quite similar in both conditions, demonstrating the capabilities of the model.

The model was further evaluated in conditions like listening to spectrally modulated sounds, band-limited noises and speech, and others, e.g., [13]–[15].

3. SOUND EXTERNALIZATION

In order to be externalized, sounds need to provide correct spectral features in the direct path [16], [17]. Externalization of anechoic sounds is not affected by broadband approximations of interaural phase differences [4] but may slightly degrade if interaural time delays (ITDs) and interaural level differences (ILDs) are inconsistent [18]. Note that while reverberation is essential to accurately estimate the distance of a sound source [19], reverberation alone is not sufficient to externalize a sound [20].

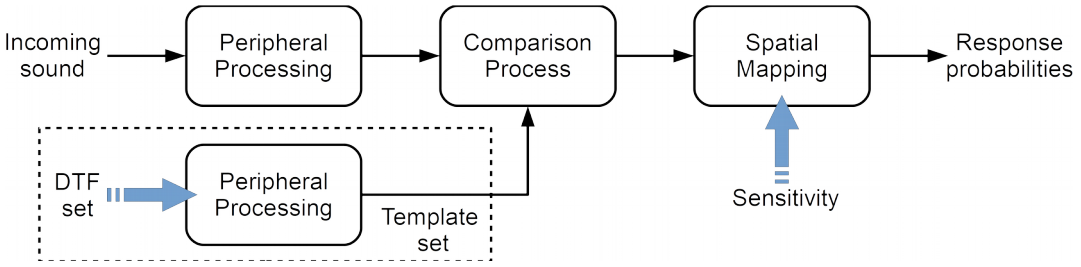


Figure 3. Structure of the sagittal-plane sound localization model from [9]. The incoming target sound is peripherally processed and the result is compared to an internal template set. The comparison result is mapped yielding the probability for responding at a given polar angle. The blue arrows indicate the free parameters of the corresponding sections, the DTF set (i.e., the directional part of HRTFs) and the sensitivity (i.e., localization precision of a listener). Figure replotted from [25].

doi:10.25836/sasp.2019.10

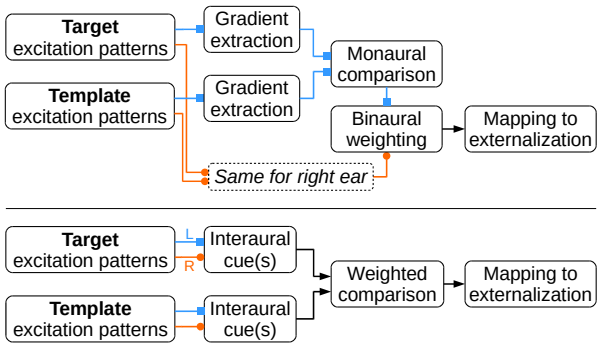


Figure 5. Structures of externalization models in our meta analysis. **Top**: Models based on monaural features. **Bottom**: Models based on interaural features.

A recent attempt to model sound externalization compared interaural spectral level differences (ISLDs) between the stimulus and internal reference templates [16]. However, interaural spectral comparisons are in contradiction to the current understanding of the monaural processing of spectral localization cues and to experiments showing that spectral manipulations degrade externalization also if ISLDs are preserved [4]. To clarify this discrepancy, we performed a model-based meta-analysis aiming at shedding light on the perceptual weighting of spectral features in comparison to other acoustic directional features.

3.1. Methods

The model architecture is based on the template matching evaluating various directional features (see Fig. 5): the positive gradients in the spectral shape of magnitude profiles (MSG) [11], the spectral shape of ILDs (ISS) [16], the spectral standard deviation of ILDs (ISSD) [21], the coherence of ITDs and ILDs (ITIC), the temporal standard deviation of ILDs (ITSD, [22], and the interaural cross-correlation (IACC) ["MaxIACCe lp" method from 23] as summarized in Tab. 1.

3.2. Results

3.2.1. Effects Of Low-frequency Alterations

In [4], the vowel */a/* was synthesized as a tone complex consisting of 38 harmonics of the fundamental frequency of 125 Hz, yielding a sound band-limited up to 4750 Hz. This sound was presented via headphones and filtered with individualized HRTFs. In one experiment, the magnitudes of all harmonics up to a certain harmonic *n'* were set to the interaural average, effectively removing ILDs up to that harmonic's frequency. In the other experiment, the ipsilateral magnitude spectrum was flattened up to *n'* while the contralateral magnitudes were shifted in parallel, effectively maintaining the original ILDs but changing the monaural spectral profiles. In both experiments, the listeners were asked to rate the degree of auditory externalization on a continuous metric scale with minimum values referring to inside-the-head localization and maximum values referring to localization at the actual loudspeaker position.

Cue	Description
MSG	Monaural spectral gradients (c.f., Baumgartner et al., 2014, [9])
ISS	Interaural spectral shape (c.f., Hassager et al., 2016, [16])
ISSD	Interaural spectral standard deviation (c.f., Georganti et al., 2013, [21])
ITIC	Interaural time-intensity trading (ITD vs. ILD)
ITSD	Interaural temporal standard deviation (c.f., Catic et al., 2015, [22])
IACC	Interaural cross-correlation (c.f., Katz and Noisternig, 2014, [23])

Table 1. Acoustic features evaluated by the externalization models.

The model predictions show that the IACC and MSG cues yielded the smallest prediction errors across both experiments, whereby there was a marked difference in predictive power between the two different experimental manipulations. The interaction term comprising MSG and ITIC performed best.

Cue	Sensitivity	Error	Weight
MSG	26.05	0.22	0.49
ISS	0.70	0.35	0.00
ISSD	1.19	0.37	0.00
ITIC	0.77	0.34	0.00
IACC	1.12	0.22	0.51
Comb.		0.15	

Table 2. Effect of low-frequency alterations up to a harmonic of a complex tone with a fundamental frequency of 125 Hz. The table shows cue-specific sensitivities for the externalization mapping (Sensitivity), prediction errors (Error), and optimal relative weights (Weight) in order to best explain the results from [4].

3.2.2. Effects Of Spectral Smoothing

In [16], Gaussian white but band-limited (from 50 to 6000 Hz) noises were filtered with individualized BRIRs (RT60 between 300 and 600 ms) in order to simulate sound sources positioned at azimuths of 0° and 50°. As independent experimental variable, Gammatone filters with various equivalent rectangular bandwidths (ERBs) were used to spectrally smooth the direct path portion (until 3.8 ms) of the BRIRs. Filters with larger ERBs more strongly smoothed the shape of the magnitude spectrum. Listeners rated auditory externalization on a continuous scale as in [4].

Model simulations were based on (anechoic) HRTFs because the original BRIRs were not accessible. This is not critical assuming that only the direct path is relevant if also only the direct path has been modified during the experiment. Despite the non-monotonicities of the ISSD cue, most predictions followed the systematic trend of externalization degradation with increasing bandwidth factor. Moreover, all these cues were consistent with the actual results in that they were insensitive to spectral smoothing below one ERB. This was particularly inter-

esting for the IACC cue, for which the model does not apply Gammatone filtering. As shown in Tab. 3, the actual results were best predicted on the basis of MSG and IACC cues.

Cue	Sensitivity	Error	Weight
MSG	24.05	0.18	0.37
ISS	2.12	0.25	0.16
ISSD	3.93	0.34	0.17
ITIC	1.75	0.28	0.00
IACC	3.79	0.22	0.29
Comb.		0.13	

Table 3. Effect of spectral smoothing on the externalization of sounds presented at the azimuth angles of 0° and 50°. Cue-specific sensitivities (Sensitivity) for the externalization mapping, prediction errors (Error), and optimal relative weights (Weight) in order to best explain the results from [16].

In [17], further tests were done for 15 listeners on the effect of spectral smoothing, focusing on the high-frequency range between 1 to 16 kHz where the pinna induces the most significant directional spectral variations. In contrast to the other studies, listeners judged auditory externalization not absolutely but relatively within paired comparisons. Absolute externalization scores were then estimated from the paired judgments via probabilistic model fitting, [24]. Scale estimates for six out of twelve listeners fulfilled the requirement of proper fitting.

Model predictions were based on listener-specific HRTFs and also assessed against listener-specific results. As shown in Tab. 4, the IACC cue yielded insufficient results (negative sensitivity and large prediction error) because the stimuli were high-pass filtered and only frequencies up to 3 kHz contribute to the IACC cue. All other cues perform quite similarly across the very limited set of experimental conditions.

Cue	Sensitivity	Error	Weight
MSG	33.46	0.35	0.34
ISS	2.51	0.33	0.00
ISSD	1.64	0.34	0.00
ITIC	2.82	0.33	0.66
IACC	-3.66	0.45	0.00
Comb.		0.32	

Table 4. Effect of spectral smoothing on the externalization of sounds two azimuths (±90° or 0°). Cue-specific sensitivities (Sensitivity) for the externalization mapping, prediction errors (Error), and optimal relative weights (Weight) in order to best explain the results from [17].

3.2.3. Effects Of BRIR Modifications And Stimulus Bandwidth

In [20], individualized BRIRs were used to simulate a talker positioned at 30° azimuth. Externalization ratings for in-the-ear (ITE) and behind-the-ear (BTE) microphone casings were collected in two conditions: broadband (BB) and 6.5-kHz-low-pass (LP) filtered speech samples, both at various mixing ratios with stereophonic recordings only providing ITD cues. The amount of reverberation remained constant across mixing ratios.

For model simulations, original BRIRs were only available for 3 out of 7 (normal-hearing) listeners. The simulation results in Tab. 5 show that the current implementation of the IACC cue clearly failed to predict the actual results from a reverberant listening condition. The MSG cue was the second worst cue mainly because it yielded predictions that overestimate the effect of low-pass filtering. All other cues performed similarly well, including the ITSD cue that has only been investigated in this reverberant setting.

Cue	Sensitivity	Error	Weight
MSG	13.00	0.22	0.05
ISS	1.44	0.17	0.14
ISSD	2.91	0.16	0.21
ITSD	3.12	0.15	0.30
ITIC	0.81	0.18	0.25
IACC	0.04	0.46	0.05
Comb.		0.13	

Table 5: Effects of BRIR modifications and stimulus bandwidth. Cue-specific sensitivities (Sensitivity) for the externalization mapping, prediction errors (Error), and optimal relative weights (Weight) in order to best explain the results from [20].

4. CONCLUSIONS

Spatial audio quality comprises many aspects of hearing. Sound localization in the sagittal planes and sound externalization are the basic qualities strongly depending on the listener-specific HRTFs. While predictions of a listener's sagittal-plane sound localization are already quite accurate, sound externalization with its variety of contributing acoustic features is far less understood.

Our meta-analysis of sound externalization studies performed under anechoic conditions showed strong indications for the relevance of monaural spectral cues (MSGs) and inconsistencies between ITD and ILD (ITICs). The overall picture suggests that listeners base their externalization ratings on one of these two cues, probably the one deviating the most from the reference sound. Under reverberant conditions, all interaural cues (ISS, ISSD, IACC, ITIC) yielded similarly small prediction errors, slightly favoring the ITIC cue. Future externalization experiments should more clearly disentangle the perceptual relevance of MSG versus ITIC features.

For both sound externalization and localization, the nature of the processing of the MSG cues by the auditory system is still poorly understood. The MSG feature implemented here has been motivated by physiological findings in cats and psychoacoustic model simulations in the context of sagittal-plane sound localization. Hence, future externalization experiments are required to unveil the details of the MSG feature, for instance by dissociating positive versus negative spectral gradients.

doi:10.25836/sasp.2019.10

5. ACKNOWLEDGMENTS

We thank Bill Withmer for kindly providing the original data from [20]. Work supported by the Austrian Science Fund (FWF J 3803-N30) and Facebook Reality Labs.

6. REFERENCES

[1] H. Møller, M. F. Sørensen, D. Hammershøi, and C. B. Jensen, 'Head-related transfer functions of human subjects', *J Audio Eng Soc*, vol. 43, pp. 300–321, May 1995.

[2] P. Majdak, P. Balazs, and B. Laback, 'Multiple exponential sweep method for fast measurement of head-related transfer functions', *J Audio Eng Soc*, vol. 55, pp. 623–637, Aug. 2007.

[3] E. H. Langendijk and A. W. Bronkhorst, 'Fidelity of three-dimensional-sound reproduction using a virtual auditory display', *J Acoust Soc Am*, vol. 107, no. 1, pp. 528–37, Jan. 2000.

[4] W. M. Hartmann and A. Wittenberg, 'On the externalization of sound images', *J Acoust Soc Am*, vol. 99, no. 6, pp. 3678–88, Jun. 1996.

[5] M. Morimoto and H. Aokata, 'Localization cues in the upper hemisphere', *J Acoust Soc Jpn E*, vol. 5, pp. 165–173, 1984.

[6] T. Okano, L. L. Beranek, and T. Hidaka, 'Relations among interaural cross-correlation coefficient (IACCE), lateral fraction (LFE), and apparent source width (ASW) in concert halls', *J Acoust Soc Am*, vol. 104, no. 1, pp. 255–65, Jul. 1998.

[7] J. S. Bradley and G. A. Soulodre, 'Objective measures of listener envelopment', *J. Acoust. Soc. Am.*, vol. 98, pp. 2590–2597, 1995.

[8] E. H. A. Langendijk and A. W. Bronkhorst, 'Contribution of spectral cues to human sound localization', *J Acoust Soc Am*, vol. 112, no. 4, pp. 1583–96, Oct. 2002.

[9] R. Baumgartner, P. Majdak, and B. Laback, 'Modeling sound-source localization in sagittal planes for human listeners', *J. Acoust. Soc. Am.*, vol. 136, pp. 791–802, 2014.

[10] L. A. J. Reiss and E. D. Young, 'Spectral edge sensitivity in neural circuits of the dorsal cochlear nucleus', *J Neurosci*, vol. 25, no. 14, pp. 3680–91, Apr. 2005.

[11] R. Baumgartner, P. Majdak, and B. Laback, 'Modeling the Effects of Sensorineural Hearing Loss on Sound Localization in the Median Plane', *Trends Hear.*, vol. 20, p. 2331216516662003, 2016.

[12] J. C. Middlebrooks, 'Virtual localization improved by scaling nonindividualized external-ear transfer functions in frequency', *J Acoust Soc Am*, vol. 106, no. 3 Pt 1, pp. 1493–1510, Sep. 1999.

[13] D. Marelli, R. Baumgartner, and P. Majdak, 'Efficient Approximation of Head-Related Transfer Functions in Subbands for Accurate Sound Localization', *IEEE Trans. Audio Speech Lang. Process.*, vol. 23, no. 7, pp. 1130–1143, Jul. 2015.

[14] R. Baumgartner and P. Majdak, 'Modeling Localization of Amplitude-Panned Virtual Sources in Sagittal Planes', *J Audio Eng Soc*, vol. 63, no. 7/8, pp. 562–569, 2015.

[15] M. Geronazzo, E. Peruch, F. Prandoni, and F. Avanzini, 'Applying a Single-Notch Metric to Image-Guided Head-Related Transfer Function Selection for Improved Vertical Localization', *J Audio Eng Soc*, vol. 67, no. 6, pp. 414–428, 2019.

[16] H. G. Hassager, F. Gran, and T. Dau, 'The role of spectral detail in the binaural transfer function on perceived externalization in a reverberant environment', *J. Acoust. Soc. Am.*, vol. 139, no. 5, pp. 2992–3000, 2016.

[17] R. Baumgartner *et al.*, 'Asymmetries in behavioral and neural responses to spectral cues demonstrate the generality of auditory looming bias', *Proc. Natl. Acad. Sci.*, vol. 114, no. 36, pp. 9743–9748, 2017.

[18] P. X. Zhang and W. M. Hartmann, 'On the ability of human listeners to distinguish between front and back', *Hear Res*, vol. 260, no. 1–2, pp. 30–46, Feb. 2010.

[19] A. J. Kolarik, B. C. J. Moore, P. Zahorik, S. Cirstea, and S. Pardhan, 'Auditory distance perception in humans: a review of cues, development, neuronal bases, and effects of sensory loss', *Atten. Percept. Psychophys.*, vol. 78, no. 2, pp. 373–395, Feb. 2016.

[20] A. W. Boyd, W. M. Whitmer, J. J. Soraghan, and M. A. Akeroyd, 'Auditory externalization in hearing-impaired listeners: The effect of pinna cues and number of talkers', *J. Acoust. Soc. Am.*, vol. 131, no. 3, pp. EL268–EL274, 2012.

[21] E. Georganti, T. May, S. van de Par, and J. Mourjopoulos, 'Sound Source Distance Estimation in Rooms based on Statistical Properties of Binaural Signals', *Audio Speech Lang. Process. IEEE Trans. On*, vol. 21, no. 8, pp. 1727–1741, Aug. 2013.

[22] J. Catic, S. Santurette, and T. Dau, 'The role of reverberation-related binaural cues in the externalization of speech', *J. Acoust. Soc. Am.*, vol. 138, no. 2, pp. 1154–1167, 2015.

[23] B. F. G. Katz and M. Noisternig, 'A comparative study of Interaural Time Delay estimation methods', *J. Acoust. Soc. Am.*, vol. 135, no. 6, pp. 3530–3540, Jun. 2014.

[24] F. Wickelmaier and C. Schmid, 'A Matlab function to estimate choice model parameters from paired-comparison data', *Behav. Res. Methods Instrum. Comput.*, vol. 36, no. 1, pp. 29–40, Feb. 2004.

[25] P. Majdak, R. Baumgartner, and B. Laback, 'Acoustic and non-acoustic factors in modeling listener-specific performance of sagittal-plane sound localization', *Front. Psychol.*, vol. 5, p. 319, 2014.

FLEXIBLE BINAURAL RESYNTHESIS OF ROOM IMPULSE RESPONSES FOR AUGMENTED REALITY RESEARCH

Sebastià V. Amengual Garí, W. Owen Brimijoin, Henrik G. Hassager, Philip W. Robinson
Facebook Reality Labs
sebastia.amengual@oculus.com

ABSTRACT

A basic building block of audio for Augmented Reality (AR) is the use of virtual sound sources layered on top of real sources present in an environment. In order to perceive these virtual sources as belonging to the natural scene it is important to carefully replicate the room acoustics of the listening space. However, it is unclear to what extent the real and virtual room impulse responses (RIR) need to be matched in order to generate plausible scenes in which virtual sound sources blend seamlessly with real sound sources. This contribution presents an auralization framework that allows binaural rendering, manipulation and reproduction of room acoustics in augmented reality scenarios, in order to get a better understanding of the perceptual relevance of individual room acoustic parameters. Auralizations are generated from measured multichannel room impulse responses (MRIR) parametrized using the Spatial Decomposition Method (SDM). An alternative method to correct the known time-dependent coloration of SDM based auralizations is presented. Instrumental validation shows that re-synthesized binaural room impulse responses (BRIRs) are in close agreement with measured BRIRs. In situ perceptual validation with expert listeners show that - in the presence of visual cues, an explicit sound reference and unlimited listening time, they are able to discriminate between a real loudspeaker and its re-synthesized version. However, the renderings appear to be as plausible as a real source once visual cues are removed. Finally, approaches to manipulate the spatial and time-energy properties of the auralizations are presented.

1. INTRODUCTION

One of the main applications of audio for AR is the integration of virtual objects into the real environment. These could be either augmented versions of real objects, e.g., adding to or changing the sound from real objects - or the creation of fully virtual objects, e.g., human avatars. In both cases, in order to deliver a convincing and coherent experience, the acoustic properties of the virtual and real environment must match. However, multiple elements are involved in this audio rendering pipeline i.e. individualization of head-related transfer functions (HRTFs), sound propagation modeling, headphone equalization, tracking latency, spatial resolution. The individual perceptual relevance of each element and appropriate error metrics are not clear.

The aim of the auralization system presented in this paper is to reproduce *in situ* a re-synthesized version of a loudspeaker in a room, over open headphones, with three degrees of rotational freedom. Such a system provides the possibility to seamlessly alternate between the real source present in the room, and the same source presented virtually over tracked headphones. We can then arbitrarily modify each of the elements of the audio rendering pipeline in order to study the effect of each component individually.

2. RE-SYNTHESIS

2.1 RIR Measurements

A single point-to-point multichannel room impulse response (MRIR) is required to re-synthesize a binaural room impulse response at a fixed location in a room. To this end, a swept sine signal is reproduced by a loudspeaker and recorded by an open microphone array. The room impulse response is obtained by deconvolving the original swept sine from the recording. The microphone array is composed of 6 miniature microphones (DPA 4060) arranged in pairs along perpendicular axes at 5 cm from an Earthworks M30 microphone at the center of the array. The holder for the microphones is 3D printed (see Fig. 1).

For validation and equalization, reference measurements are obtained at the receiver position using a binaural mannequin (G.R.A.S. KEMAR).

2.2 Spatial analysis

The analysis of the MRIR is based on the Spatial Decomposition Method (SDM) [1], and is implemented using the SDM Toolbox [7]. The concept behind SDM is based on the assumption that the measured soundfield can be represented as a succession of discrete acoustic events, each of them represented by a sample in the RIR with an associated direction-of-arrival (DOA). This assumption is certainly violated after more than one acoustic event arrive simultaneously at the receiver location, due to increasing

© Sebastià V. Amengual Garí, W. Owen Brimijoin, Henrik G. Hassager, Philip W. Robinson. Licensed under a Creative Commons Attribution 4.0 International License (CC BY 4.0). **Attribution:** Sebastià V. Amengual Garí, W. Owen Brimijoin, Henrik G. Hassager, Philip W. Robinson. "Flexible binaural resynthesis of room impulse responses for augmented reality research", 1st EAA Spatial Audio Signal Processing Symposium, Paris, France, 2019.

doi:10.25836/sasp.2019.31



Figure 1. Custom microphone array (5 cm radius).

echo density. However, increasing echo density eventually results in a diffuse sound field, which can be regarded as a succession of events with random DOA. The method has been successfully used to implement analysis and auralization of sound fields in a variety of scenarios, including concert halls [2], stage acoustics [3], or car cabin acoustics [4, 5], among others. In the present study, we use the broadband RIRs from the 7 microphones as input for the direction-of-arrival (DOA) estimation. The SDM analysis window is set to the smallest allowed size, and the resulting DOA vectors are smoothed using a moving average window of 16 samples (with a sampling rate of 48 kHz).

2.3 BRIR Synthesis

The SDM sound field parametrization consists of a [1×N] vector $\mathbf{p} = [p_1, p_2, ..., p_N]$ containing the pressure RIR and a [3×N] matrix $\mathbf{r} = [\mathbf{r}_1, \mathbf{r}_2, ..., \mathbf{r}_N]$ indicating the DOA for each of the samples in cartesian coordinates. A binaural reconstruction can be implemented in a straightforward manner by combining the DOA data and the pressure RIR with a Head-Related Impulse Response (HRIR) dataset. A binaural room impulse response (BRIR) can be reconstructed as a weighted summation of delayed HRIRs corresponding to the closest directions to each sound event. A set of BRIRs corresponding to arbitrary head rotations, can be created by rotating the DOA matrix.

At first, the indices $\hat{k}_n^u$ to the relevant HRIRs for each sound event in each head orientation, u, are found. This is done by finding the nearest HRIR for each sample, n, in the rotated DOA matrices $\mathbf{r}^u$.

$$\hat{k}_n^u = \underset{k \in 1,...,K}{\arg\min} \{ \mathrm{d}(r_n^u, \hat{\mathbf{r}}) \} \tag{1}$$

where $\hat{\mathbf{r}}$ is a $[3 \times K]$ matrix containing the source/receiver relative orientations of the HRIR dataset in cartesian coordinates and $\mathrm{d}(\cdot, \cdot)$ is the Euclidean distance. The rotated DOA matrices are created by multiplying $\mathbf{r}$ with corresponding rotation matrices.

$$\mathbf{r}^u = R_z(-\theta^u)\, R_y(-\phi^u)\, \mathbf{r} \tag{2}$$

where R_y and R_z rotation matrices to render an arbitrary head orientation (θ^u, ϕ^u). This applies to a right-hand rule coordinate system where positive Y is left and positive Z is up [1] .

[1] To achieve a correct head rotation the DOA has to be rotated in a reversed order. Roll rotation is excluded from the equation as it is not implemented in the framework.

Next, the BRIRs for all head orientations can be constructed using the indices $\hat{k}_n^u$ of the HRIRs, by delaying the HRIR at the nth positions by n samples and multiplying it by the instantaneous pressure, p_n:

$$\mathbf{BRIR}^u(t) = \sum_{n=1}^{N} p_n\, \mathbf{HRIR}_{\hat{k}_n^u} \circledast \delta(t - n), \tag{3}$$

where **HRIR** is a three-dimensional $[H \times K \times 2]$ matrix containing a HRIR dataset of H samples (per channel) and K source/receiver relative orientations, and t indicates the samples in the BRIR.

2.4 Reverb correction

A known artifact of SDM based auralizations is the increase of high frequency energy in the late reverb, leading to a whitening of the spectrum. A description of this problem and a time-frequency equalization approach was presented in [4]. This technique is especially useful for loudspeaker based auralizations and binaural renderings based on a virtual loudspeaker approach, as it uses the pressure RIR $\mathbf{p}$ as a reference. However, if a HRIR dataset with high spatial resolution is used - hence, the virtual loudspeaker set-up contains many virtual loudspeakers - the time-frequency filtering becomes time and resource consuming, as each separate stream needs to be equalized. For this reason, an alternative approach to compensate for the late reverberation coloration is presented below.

The main advantage of this approach is that only one reverberation equalization needs to be performed per head orientation, as it is performed on the re-synthesized BRIR. Two main assumptions are made here: 1) the reverberation time (RT_{60}) of all the re-synthesized BRIRs must be the same as the one from the pressure RIR, regardless of the head orientation and 2) the time-frequency deviations of the reverberation in the synthesized BRIRs are not time dependent. It is expected that these assumptions hold true in most situations and could be only violated in extreme cases of highly directional late reverberation or double decay slopes.

The first step is to decompose the pressure RIR, $\mathbf{p}$, and the re-synthesized BRIR, $\mathbf{BRIR}^u$ in fractional octave bands by using perfect reconstruction filters [6]. The implementation of the filters is largely based on that present in the SDM Toolbox [7]. Then, the frequency dependent RT_{60} of the pressure RIR, $\mathbf{RT}_{60,\mathrm{orig}}$ and the RT_{60} of the re-synthesized BRIR, $\mathbf{RT}_{60,\mathrm{resynth}}$ are computed. The computed frequency dependent RT_{60} is used to generate the parameters of an exponential function that is multiplied with each band of the re-synthesized BRIR to modify their reverberation time, following the method first presented in [8]. Finally, all of the corrected subbands $\mathbf{BRIR}^u_{\mathrm{corr},f}$ are summed together resulting in the corrected BRIR, $\mathbf{BRIR}^u_{\mathrm{corr}}$.

$$\mathbf{BRIR}^u_{\mathrm{corr}}(t) = \sum_{f=1}^{F} \mathbf{BRIR}_{\mathrm{corr},f}(t) \tag{4}$$

$$\mathbf{BRIR}^u_{\mathrm{corr},f}(t) = \mathbf{BRIR}^u_f(t) e^{-t(d_{1,f} - d_{0,f})} \tag{5}$$

doi:10.25836/sasp.2019.31

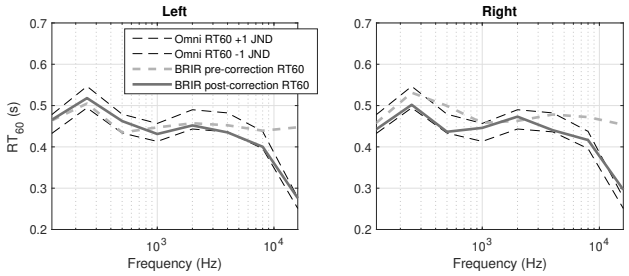


Figure 2. Comparison of the original RT_{60} and the re-synthesis before and after correction.

The constants of the exponential function, $d_{1,f}$ and $d_{0,f}$, are derived from the RT_{60} as follows

$$d_{0,f} = \frac{\ln(10^6)}{2\,RT_{60,\text{resynth},f}} \tag{6}$$

$$d_{1,f} = \frac{\ln(10^6)}{2\,RT_{60,\text{orig},f}} \tag{7}$$

A comparison between the estimated reverberation time of the original pressure RIR and the re-synthesized BRIRs is presented in Fig. 2. It is seen that before the correction the reverberation time is overestimated at high frequencies. After correction, the reverberation time falls within ± 1 JND (5% of the RT60, as defined in the ISO 3382).

2.5 Monaural equalization

The re-synthesized BRIRs can be regarded as a weighted summation of delayed HRTFs. Ideally, the spectral properties of the HRTF used for the re-synthesis should match perfectly with those of the subject listening to the re-synthesized BRIRs. However, the acquisition method of the HRTFs does in many cases introduce spectral distortions that are not easily characterized. For instance, in the case of simulated HRTFs, the skin absorption properties must be modeled, or in certain HRTF measurement systems the low frequency response must be extrapolated. For this reason, we use a mannequin measurement conducted at the listening position to generate a direction independent monaural filter that is applied to all the BRIRs at the reproduction stage. This equalization accounts as well for frequency response deviations of the omnidirectional RIR.

The filter is derived by dividing the smoothed magnitude response (1/12 octave band) of the left and right channels of a measured BRIR and its re-synthesis, ensuring that the relative source/receiver orientation is the same. The magnitude differences are averaged over left and right channels and a minimum phase FIR filter of 2048 taps is generated. Design of the filter is done using the `fdesign` and `rceps` functions of Matlab.

During experimental validation we found that the differences between equalization filters derived from various relative source/receiver orientations are negligible below approximately 12 kHz. At high frequencies, HRTF datasets are prone to present artifacts due to simulation limitations or spurious reflections in measurements. Thus we can consider this filter to be direction independent and the equalization is applied for any arbitrary orientation. Fig. 3 shows

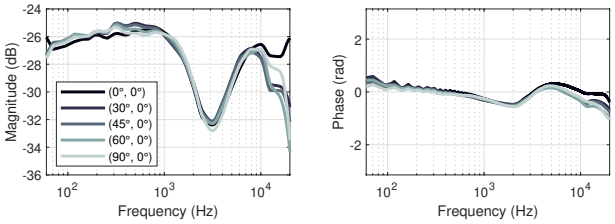


Figure 3. Magnitude and phase responses of a monaural equalization filter computed for various head orientations.

the monaural equalization filter response for various head orientations. In this case, the HRTF used for re-synthesis was generated from simulations, and thus a generalized spectral mismatch was expected.

3. OBJECTIVE VALIDATION

An objective validation comparing measured BRIRs and their re-synthesized counterparts is presented in Fig. 4. Although the re-synthesis method does not aim at a physical reconstruction of the BRIRs, a close agreement is found between the original and the re-synthesized BRIRs. The results are presented for two rooms with different source/receiver orientations (frontal orientation in room A, and lateral orientation in room B).

The re-synthesized broadband pressure BRIRs resembles the overall time-energy structure of the original BRIRs, correctly displaying appropriate timing and amplitude of the most energetic events. However, some spurious reflections can be observed in the re-synthesis, most likely due to constructive interference of acoustic events when combining HRIRs for various directions.

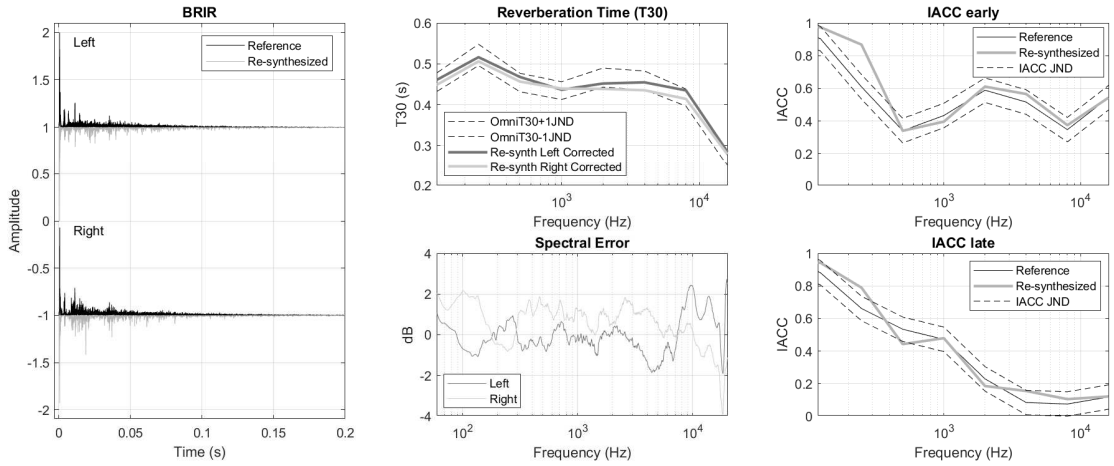
The RT_{60} of the re-synthesis fits within ± 1 JND (5% of the reference RT_{60}, as defined in ISO 3382) at most of the frequency bands.

The spectral error after monaural equalization falls within ± 2 dB at all frequencies up to 10 kHz. More variability is encountered between 10 and 20 kHz, likely due to uncertainties in the used HRTFs. The HRTF dataset used in this validation is simulated using the boundary element method (BEM) and a 3D mesh corresponding to the same mannequin used in the BRIR measurements (G.R.A.S. KEMAR).

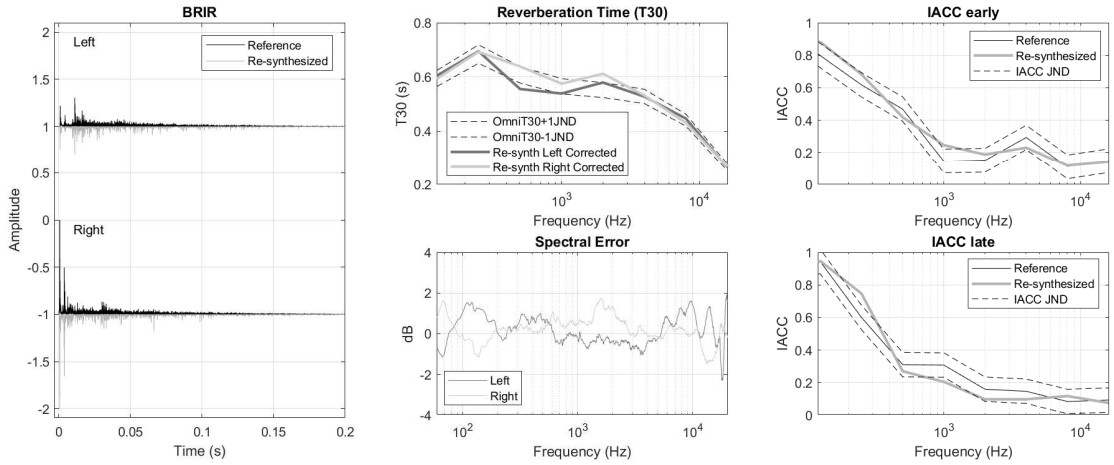
The early and late interaural cross correlation are generally close to ± 1 JND (0.075, as defined in ISO 3382) at all frequency bands, suggesting that the spatial properties of the re-synthesized BRIRs largely resemble those of the original BRIRs. Note that the late reverberation tails used in the re-synthesized BRIRs correspond to an arbitrary direction, different than those of the early reflections. This supports the assumption that the late reverberation can effectively be rendered as a direction independent filter [9].

4. REAL-TIME FRAMEWORK

An experimental framework has been developed in Max/MSP to handle the convolution of the BRIRs with an



a) Room A (frontal position).



b) Room B (lateral position).

Figure 4. Instrumental validation of the re-synthesized BRIRs as compared to the in-situ mannequin measurements.

anechoic signal, headphone equalization, monaural HRTF frontal/diffuse equalization and listener tracking. The framework can be interfaced with other Max/MSP patches allowing quick prototyping of listening tests by switching BRIR datasets and equalization filters seamlessly. The (end-to-end) motion to sound latency of the system is approximately 70 ms.

4.1 Convolution

The real-time operations for dynamic rendering are performed in Max/MSP. The convolution pipeline is split in 3 separate parts: direct sound, early reflections, and late reverberation. The direct sound (first 128 samples) and the early reflections are convolved dynamically, and filters are switched as head movements are detected. An equal power cross-fade with a duration of 1 ms is applied when switching filters to avoid audible artifacts. The convolution operations are performed using the spat5.conv~ object.

If the mixing time is sufficiently high, the late reverberation can be modeled as a direction independent filter [9]. The framework implements a single binaural FIR convolution and the mixing time between early reflections and late reverberation can be arbitrarily adjusted. A window of variable size is implemented to ensure a smooth transition between the early and late part of the BRIR.

4.2 Tracking

The framework handles real-time tracking of the listener orientation and loudspeaker positions, ensuring that the relative angles between the listener and the real and virtual sources match. The loudspeaker tracking is implemented using an optical tracking system (OptiTrack or Vicon), while listener tracking is done either with an optical system or an Oculus Rift headset.

5. PERCEPTUAL VALIDATION

A pilot study to assess the authenticity of the re-synthesized BRIRs has been conducted, and a second test assessing plausibility is currently in progress. The authenticity test was composed of three parts: a discrimination test (ABX), an identification test (2 alternative forced choice - 2AFC) and a qualitative rating. Ten expert listeners wearing non-occluding headphones (AKG K1000) were presented with sounds coming from a loudspeaker,

doi:10.25836/sasp.2019.31

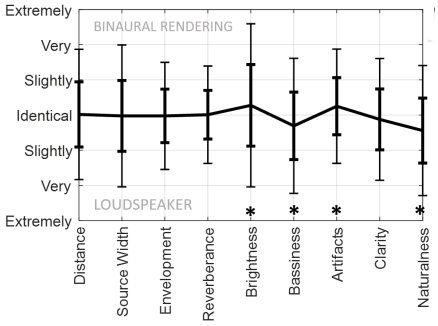


Figure 5. Perceptual ratings of the binaural renderings as compared to a real loudspeaker.

and a re-synthesized version using individualized HRTFs and headphone equalization. They were allowed to switch between presentation modes in real time and were granted unlimited listening time. In addition, the loudspeakers were visible. Three source positions and four sound samples were used (male and female speech, solo guitar, and pink noise), resulting in 12 trials. In these conditions, listeners were able to correctly discriminate in 114/120 (95%) of the trials, and to correctly identify which sample was the loudspeaker in 96/120 (80%) of the trials. However, only ratings of timbral cues, *naturalness* and *artifacts* presented statistically significant differences between the real loudspeaker and the renderings. In addition, several participants mentioned that in some cases a subtle localization mismatch between real and virtual sources was enough to inform their judgment in identifying which sample was the reference, and whether it was a real loudspeaker or a binaural rendering. A summary of the test results is presented in Fig. 5. It is worth noting that the participants in this study were highly familiar with binaural audio rendering and critical listening and lower discrimination and identification rates could be expected with naive listeners. To our knowledge, only one study assessed authenticity of individualized dynamic binaural audio, and although the BRIRs were generated using binaural microphones in the subjects' ears, authenticity was not achieved [10].

A pilot listening test was conducted to assess the plausibility of generic binaural renderings. In this case, a loudspeaker was hidden behind an acoustically transparent curtain and compared against a binaural re-synthesis generated with BEM simulated KEMAR HRTFs and generic headphone equalization (performed on the same mannequin). Seven expert listeners were granted unlimited time to listen to the two stimuli (castanets of approximately 7 seconds of duration) and were able to switch between them in real time. Then, they were asked which of the two sounds was perceived as more plausible - or likely to be a loudspeaker. Preliminary results suggest that in the absence of visual cues, and even with generic HRTFs, the perceived plausibility of virtual sounds is as high as that perceived for a real loudspeaker. At the moment of writing this manuscript a full test is being conducted to confirm these initial results.

6. BRIR MANIPULATIONS

The premise behind the implementation of this auralization system is to enable perceptual research in room acoustics. This section presents a variety of manipulations that we implemented in order to test the perceptual effects of various parts of the rendering pipeline.

6.1 Pre-rendered manipulations

A number of straightforward modifications in the re-synthesis algorithm allow the rendering of time-energy and spatial manipulations on the generated BRIRs. Those manipulations that require the modification of the HRTF dataset, the pressure RIR or the DOA vectors must generally be performed offline, prior to the rendering of re-synthesized BRIRs.

6.1.1 HRTF Dataset

The re-synthesis minimizes the distance between the DOA of an acoustic event and the HRIR used to render such event. Arbitrary manipulations of the spatial grid of the HRTF can be used to investigate the requirements of spatial resolution of HRTFs. Minimizing the size of a HRTF dataset to decrease memory requirements is important in mobile systems rendering spatial audio in real time.

6.1.2 DOA Quantization

The DOA vectors generated by SDM can be directly manipulated to render spatial manipulations of the sound field. For instance, quantizing the DOA vectors in order to limit the possible DOA of reflections while keeping intact the DOA of the direct sound can be useful to investigate the spatial resolution needed for the reproduction of reflections. In real-time systems, minimizing the number of reflection directions can contribute to decreasing the number of convolutions, optimizing compute requirements.

6.1.3 Reverberation time

The strategy introduced in [8] and used in Section 2.4 to correct the RT_{60} of re-synthesized BRIRs can be applied to implement arbitrary manipulations of the reverberation time. For example, modifying the reverberation time of a rendered BRIR and performing in-situ comparisons against a real loudspeaker allows the study of perceptual thresholds of the room acoustic divergence effect [11].

6.1.4 Synthetic reverberation

The late reverberation tail is rendered as a direction independent filter and it can be easily swapped with a synthetic reverberation. For instance, it is common for real-time sound engines to use reverberators for the synthesis of late reverberation. By combining spatial data with a synthetic monaural reverberation tail, various architectures can be compared.

6.2 Real-time manipulations

Those manipulations that are related to the real-time convolution of the BRIRs with anechoic audio can be rendered

in real-time, as they not require the modification of any data prior to the re-synthesis of the BRIRs.

6.2.1 Relative levels of the RIR

Another manipulation related to the study of the room divergence effect [11] is that of the relative levels between the direct sound, early reflections and late reverberation. As the re-synthesized BRIRs are convolved in three separate pipelines, it is straightforward to modify the relative levels of each part of the BRIR. This is especially useful to study the link between the direct-to-reverberant ratio (DRR) and the perception of distance.

6.2.2 Direction dependent reverberation level

In [12], the importance of reverberation at the ipsi- and contralateral ear and their role in perceived externalization was investigated. However, the study was limited to static sources. The rendering approach presented here allows for the implementation of dynamic and direction dependent modifications of the reverberation level at each ear.

6.2.3 Distortion of the acoustic space

In anechoic conditions, the perceived angle of a sound source suffers angle-dependent distortions [13], i.e., the perceived direction does not correspond to the actual direction of reproduction. A module to compensate for spatial distortions is included to enable further investigation in reverberant conditions.

7. SUMMARY AND CONCLUSIONS

A framework for the re-synthesis of BRIRs based on SDM analysis is presented. An alternative approach to correct the known reverberation artifacts of SDM auralizations is introduced and validated. The auralizations can be dynamically reproduced using head tracking in a Max/MSP framework. An objective validation of the re-synthesized BRIRs by directly comparing them to measured BRIRs suggests that the time-energy, spatial and spectral properties are largely replicated, with minor deviations. The renders are found to be perceptually plausible when compared to real loudspeakers, although not indistinguishable from an explicit reference. Finally, a variety of manipulations of the re-synthesized BRIRs are introduced, enabling the study of BRIR degradations in direct comparison to real loudspeakers. Future work will focus on the study of perceptual thresholds in the degradation of BRIRs in order to establish the extent of acceptable acoustic deviations in the rendering of virtual sounds in augmented reality.

8. REFERENCES

[1] S. Tervo, J. Pätynen, A. Kuusinen, and T. Lokki, "Spatial decomposition method for room impulse responses," *J. Audio Eng. Soc*, vol. 61, no. 1/2, pp. 17–28, 2013.

[2] J. Pätynen, S. Tervo, and T. Lokki, "Analysis of concert hall acoustics via visualizations of time-frequency and spatiotemporal responses," *The Journal of the Acoustical Society of America*, vol. 133, no. 2, pp. 842–857, 2013.

[3] S. Amengual Garí, M. Kob, and T. Lokki, "Investigations on stage acoustic preferences of solo trumpet players using virtual acoustics," Proceedings of the 14th Sound and Music Computing Conference 2017, pp. 8; 167–174, 2017-07.

[4] S. Tervo, J. Pätynen, N. Kaplanis, M. Lydolf, S. Bech, and T. Lokki, "Spatial analysis and synthesis of car audio system and car cabin acoustics with a compact microphone array," *J. Audio Eng. Soc*, vol. 63, no. 11, pp. 914–925, 2015.

[5] N. Kaplanis, S. Bech, S. Tervo, J. Pätynen, T. Lokki, T. van Waterschoot, and S. H. Jensen, "Perceptual aspects of reproduced sound in car cabin acoustics," *The Journal of the Acoustical Society of America*, vol. 141, no. 3, pp. 1459–1469, 2017.

[6] J. Antoni, "Orthogonal-like fractional-octave-band filters," *The Journal of the Acoustical Society of America*, vol. 127, no. 2, pp. 884–895, 2010.

[7] S. Tervo, "SDM Toolbox." `https://www.mathworks.com/matlabcentral/fileexchange/56663-sdm-toolbox`. Accessed: 2019-06-20.

[8] D. Cabrera, D. Lee, M. Yadav, and W. L. Martens, "Decay envelope manipulation of room impulse responses: Techniques for auralization and sonification," in *Proceedings of Acoustics 2011*, (Gold Coast, Australia), 2011.

[9] A. Lindau, L. Kosanke, and S. Weinzierl, "Perceptual evaluation of model- and signal-based predictors of the mixing time in binaural room impulse responses," *J. Audio Eng. Soc*, vol. 60, no. 11, pp. 887–898, 2012.

[10] F. Brinkmann, A. Lindau, and S. Weinzierl, "On the authenticity of individual dynamic binaural synthesis," *The Journal of the Acoustical Society of America*, vol. 142, no. 4, pp. 1784–1795, 2017.

[11] S. Werner, F. Klein, T. Mayenfels, and K. Brandenburg, "A summary on acoustic room divergence and its effect on externalization of auditory events," in *2016 Eighth International Conference on Quality of Multimedia Experience (QoMEX)*, pp. 1–6, June 2016.

[12] S. Li, R. Schlieper, and J. Peissig, "The effect of variation of reverberation parameters in contralateral versus ipsilateral ear signals on perceived externalization of a lateral sound source in a listening room," *The Journal of the Acoustical Society of America*, vol. 144, no. 2, pp. 966–980, 2018.

[13] W. O. Brimijoin, "Angle-dependent distortions in the perceptual topology of acoustic space," *Trends in Hearing*, vol. 22, pp. 1–11, 2018.

doi:10.25836/sasp.2019.31

MULTI-ZONE SOUND FIELD REPRODUCTION VIA SPARSE SPHERICAL HARMONIC EXPANSION

Ajay Dagar and Rajesh M Hegde
Department of Electrical Engineering
Indian Institute of Technology Kanpur, India
{adagar,rhegde}@iitk.ac.in

ABSTRACT

Multi-zone sound field reproduction is a method that can create personal audio zones to individual listeners within an enclosed environment. The accuracy requirements at the listening positions of this method results in increased number of loudspeakers and the array effort used in the computation. In this paper, we develop a sparsity based framework that provides accurate 3-D multi-zone sound field reproduction using a sparse spherical harmonics expansion framework. The sparse framework provides an optimal set of minimally distributed loudspeaker positions sufficient enough to span all the directions of interest. In this work, multi-zone sound field reproduction is formulated as a non-convex optimization problem that finds the loudspeaker weights to accurately reproduce desired sound field over the bright zones and minimizes the energy flow in the dark zones. This multi objective problem is solved using a sparse iterative method based on Bregman Iteration. Sound field reconstruction error analysis along with the objective and subjective evaluation experiments are conducted to demonstrate the effectiveness of the proposed reproduction method in creating personal sound zones.

1. INTRODUCTION

Sound field reproduction with loudspeaker arrays to provide a private listening experience at multiple zones without inter-zone interference in an enclosed environment is referred as *Personal Audio*. At various places, such as exhibition centres, private vehicles, music stores and shared offices, the personalized sound zones are now becoming popular over the traditional headphone based listening [1]. In recent years, the increasing demand of personalized sound zones led to an extensive research in the area of multi-zone soundfield reproduction. Initially, the approaches [2] - [3], were based on the acoustic contrast control (ACC) where idea was to manipulate the energy flow by maximizing the sound level difference in the bright and the dark zones. In [4], authors were able to realize 19dB and 15dB of acoustic contrast under ideal and real conditions in a car environment. To further improve the robustness and enhance the spatial quality of the reproduced sound field, a modified ACC is proposed in [5]-[6]. Later in [7] - [9], the multi-zone methods rely on the pressure matching (PM) with aim to minimize the average error between the reproduced and desired sound fields at the target zones. In [7], the error minimization problem is based on least-squares (LS) approach to find the optimal loudspeaker gains. A modified LS with energy constraints and Tikhonov-regularization are also proposed in [8]- [9]. Further, a combination of ACC and PM approach can be found in [10] - [11]. Lastly, the mode matching methods in cylindrical and spherical harmonics domain to create personal sound zones are explored in [12]- [13]. In these approaches, for accurate sound field reproduction, the required active number of loudspeakers is large for a larger region of reproduction. This result in the requirement of a dense loudspeaker grid and increased array effort putting a limit on practical implementations. Thus, it is required that the methods used for multi-zone reproduction must be capable of providing the desired sound fields using an optimal set of loudspeakers having reduced array size.

In this paper, a multi-zone reproduction problem in spherical harmonics domain with sparsity constraint is proposed for accurate sound field reproduction. The method obtains an optimal set of minimally distributed loudspeaker positions from a set of all possible loudspeaker positions sufficient to reproduce the desired sound fields. This optimal set of loudspeakers has advantage in terms of reduced reproduction error when compared to the set of loudspeakers having the same array size that are chosen randomly. Similar to [14], the multi-zone reproduction is addressed as a problem of finding the global soundfield coefficient corresponding to the target pressures defined by local spherical harmonic coefficient for each zone. Later, the loudspeaker weights are obtained by solving a non-convex optimization problem with additional constraint to control the array effort required to drive the loudspeakers. The solution of the proposed method is based on the Bregman Iteration [15].

The rest of paper is organised as follows: In Section-II, system model for multi-zone soundfield reproduction is discussed. Problem formulation and sparsity based framework is presented in Section-III. Later in Section-IV, the proposed method is evaluated objectively and subjectively followed by the conclusion and future work in Section-V.



© Ajay Dagar and Rajesh M Hegde. Licensed under a Creative Commons Attribution 4.0 International License (CC BY 4.0). **Attribution:** Ajay Dagar and Rajesh M Hegde. "Multi-Zone Sound Field Reproduction Via Sparse Spherical Harmonic Expansion", 1st EAA Spatial Audio Signal Processing Symposium, Paris, France, 2019.

doi:10.25836/sasp.2019.13

2. MULTIZONE SOUND FIELD REPRODUCTION VIA SPHERICAL HARMONIC EXPANSION

Consider a multi-zone environment with Q nonoverlapping spherical zones of radius R_q located at the local origin O_q with polar coordinates (r_q, Θ_q) measured from the global origin O. Let, the zones are enclosed within a global spherical region of soundfield reproduction having radius R_g such that $R_g \geq (R_q + r_q), \forall q$. An array of L loudspeakers is assumed to be placed on a sphere of radius R_l, where $R_l > R_g$, with each loudspeaker emitting a normalised plane wave having frequency f in a direction $\Theta_l \triangleq (\theta_l, \phi_l)$, where θ and ϕ are spherical elevation and azimuth angles, respectively.

The reproduced sound field at 1^{th} point $\mathbf{r_q^1} = (r_q^1, \Theta_q^1)$ on the surface of q^{th} zone can be expressed as

$$p(k, \mathbf{r_q^1}) = \sum_{n=0}^{\infty} \sum_{m=-n}^{n} X_n(kr_q^1) \sum_{l=1}^{L} p_{nm}(k, \Theta_l) Y_n^m(\Theta_q^1) \tag{1}$$

where $X_n(kr_q^i) = 4\pi(-i)k\mathcal{H}_n(kR_l)\mathcal{J}_n(kr_q^1)$ and $p_{nm}(k, \Theta_l) = a(k, \Theta_l)[Y_n^m(\Theta_l)]^*$ with $\mathcal{J}_n(.)$ and $\mathcal{H}_n(.)$ being the n^{th} order Bessel function of first kind and Hankel function of second kind, respectively. The spherical harmonics function, $Y_n^m(\Theta)$, is given by

$$Y_n^m(\Theta) = \frac{1}{2}\sqrt{\frac{(2n+1)(n-m)!}{\pi(n+m)!}} P_n^m(\cos\theta) e^{im\phi} \tag{2}$$

where, $P_n^m(.)$ is the legendre polynomial of order n and degree m. As per Theorem-2 mentioned in [16], for an accurate soundfield reproduction, within the considered global spherical region, it is required to truncate the order n in equation (1) to $N_g = \lceil kR_g \rceil$. In matrix form, we can re-write equation (1) as

$$p(k, \mathbf{r_q^1}) = \mathbf{x}(k, \mathbf{r_q^1}) \mathbf{D}(\mathbf{\Theta}_L) \mathbf{a}(k, \mathbf{\Theta}_L) \tag{3}$$

where, $\mathbf{x}(k, \mathbf{r_q^1})$ is a vector of size $1 \times (N_g + 1)^2$ and $\mathbf{D}(\mathbf{\Theta}_L)$ is a matrix of size $N_g + 1)^2 \times L$ defined as

$$\begin{aligned}
\mathbf{x}(k, \mathbf{r_q^1}) &= \begin{bmatrix} X_0(kr_q^1)\mathbf{y}_0(\Theta_q^1) & \cdots & X_{N_g}(kr_q^1)\mathbf{y}_{N_g}(\Theta_q^1) \end{bmatrix} \\
\mathbf{D}(\mathbf{\Theta}_L) &= \begin{bmatrix} \mathbf{y}^*(\Theta_1) & \mathbf{y}^*(\Theta_2) & \cdots & \mathbf{y}^*(\Theta_L) \end{bmatrix} \\
\mathbf{y}(\Theta_l) &= \begin{bmatrix} \mathbf{y}_0(\Theta_l) & \mathbf{y}_1(\Theta_l) & \cdots & \mathbf{y}_{N_g}(\Theta_l) \end{bmatrix}^T \\
\mathbf{y}_n(\Theta) &= \begin{bmatrix} Y_n^{-n}(\Theta) & \cdots & Y_n^n(\Theta) \end{bmatrix}, \quad \forall n = 0, \ldots, N_g
\end{aligned}$$

Also, $\mathbf{a}(k, \mathbf{\Theta}_L) \in \mathbb{C}^{L \times 1}$ is the vector containing loudspeaker weights given by

$$\mathbf{a}(k, \mathbf{\Theta}_L) = \begin{bmatrix} a_k(\Theta_1) & a_k(\Theta_2) & \cdots & a_k(\Theta_L) \end{bmatrix}^T \tag{4}$$

Now, the reproduced soundfield on the surface of q^{th} zone, having I_q sample points over the sphere, $\mathbf{p}_q(k) \in \mathbb{C}^{I_q \times 1}$ is expressed as

$$\begin{aligned}
\mathbf{p}_q(k) &= \mathbf{X}_q(k) \mathbf{D}(\mathbf{\Theta}_L) \mathbf{a}(k, \mathbf{\Theta}_L) \\
&= \begin{bmatrix} p_q(k, \mathbf{r_q^1}) & p_q(k, \mathbf{r_q^2}) & \cdots & p_q(k, \mathbf{r_q^{I_q}}) \end{bmatrix}^T
\end{aligned} \tag{5}$$

where, the matrix $\mathbf{X}_q(k) = \begin{bmatrix} \mathbf{x}(k, \mathbf{r_q^1}) & \cdots & \mathbf{x}(k, \mathbf{r_q^{I_q}}) \end{bmatrix}^T$.

Now, the objective here is to find these loudspeaker gains, $\mathbf{a}(k, \mathbf{\Theta}_L)$ to reproduce the target sound field in the zone of interest and simultaneously reduce the active number of loudspeakers. The problem formulation and the sparsity framework developed to achieve the above objectives are discussed in following section.

3. TARGET SOUND FIELD REPRODUCTION VIA SPARSE SPHERICAL HARMONIC EXPANSION

Here, the target sound field model in spherical harmonic domain is discussed first. Later, the problem of multi-zone soundfield reproduction via sparse spherical harmonic expansion is formulated.

3.1 Modelling of Target Sound Fields

In this multi-zone environment, the target sound fields are defined by considering each spherical zone, either a bright zone or a dark zone, separately in terms of spherical harmonics determined with respect to the local origins, O_q.

For a bright zone, say q^{th} zone, we consider a normalised plane wave of frequency $f (k = \frac{2\pi f}{c})$ from direction Ψ_q to be reproduced at I_q sample points over the sphere. Now, the target pressure at the 1^{th} point on q^{th} zone is given as [17]

$$p_q^t(k, \mathbf{r_q^1}) = \sum_{n=0}^{\infty} \sum_{m=-n}^{n} 4\pi i^n \mathcal{J}_n(kr_q^1)[Y_n^m(\Psi_q)]^* Y_n^m(\Theta_q^1) \tag{6}$$

Since $Y_n^m(\Theta)$ are defined with respect to O_q, the order in equation (6) is truncated to $N_q = \lceil kR_q \rceil$. Now, the target pressures $\mathbf{p}_q^t(k) \in \mathbb{C}^{I_q \times 1}$ for q^{th} zone, in matrix form, are given as

$$\mathbf{p}_q^t(k) = \begin{bmatrix} p_q^t(k, \mathbf{r_q^1}) & p_q^t(k, \mathbf{r_q^2}) & \cdots & p_q^t(k, \mathbf{r_q^{I_q}}) \end{bmatrix}^T \tag{7}$$

Considering the dark zones to be complete silent region, the target pressures $p_q^t(k, \mathbf{r_q^1})$ at I_q sample points for q^{th} zone can be assumed to be zero. Depending on whether q^{th} zone is considered as bright or dark, $\mathbf{p}_q^t(k)$ will have pressures as defined in (6) or a zero vector.

3.2 Problem Formulation

The multi-zone reproduction problem is formulated by minimizing the reproduction error and finding the corresponding loudspeaker gains. Thus, the optimization problem can be formed using equation (5) and (7) as

$$\begin{aligned}
\mathbf{a}^* = \min_{\mathbf{a}(k, \mathbf{\Theta}_L)} \quad & \tfrac{1}{2} \textstyle\sum_{q=1}^{Q} \|\mathbf{p}_q(k) - \mathbf{p}_q^t(k)\|_2^2 \\
\text{s.t.} \quad & \mathbf{p}_q(k) = \mathbf{X}_q(k) \mathbf{D}(\mathbf{\Theta}_L) \mathbf{a}(k, \mathbf{\Theta}_L), \forall q \\
& \mathbf{a}^T(k, \mathbf{\Theta}_L) \mathbf{a}(k, \mathbf{\Theta}_L) \leq \beta
\end{aligned} \tag{8}$$

where, the second constraint on $\mathbf{a}(k, \mathbf{\Theta}_L)$ ensures that array effort is maintained below its maximum value β. Apparently when L is large, the weight vector $\mathbf{a}(k, \mathbf{\Theta}_L)$ will be sparse corresponding to lesser number of active loudspeakers. As a result, $\mathbf{D}(\mathbf{\Theta}_L)$ containing the spatial information will also become sparse in spatial domain.

doi:10.25836/sasp.2019.13

Considering an overcomplete dictionary matrix containing spherical harmonic functions, $\mathbf{S}(\mathbf{\Theta}_\tau) \in \mathbb{C}^{(N_g+1)^2 \times \tau}$ and a sparse matrix, $\mathbf{R} \in \mathbb{R}^{\tau \times L}$ with $\left[\mathbf{R}\right]_{ij} \in \{0,1\}$, we can write

$$\mathbf{D}(\mathbf{\Theta}_L) = \mathbf{S}(\mathbf{\Theta}_\tau)\mathbf{R} \tag{9}$$

where, $\mathbf{D}^T(\mathbf{\Theta}_L)\mathbf{D}(\mathbf{\Theta}_L) = \mathbf{I}$ as spherical harmonics functions forms the matrix $\mathbf{D}(\mathbf{\Theta}_L)$. Thus, the expression in (9) is valid only when the orthonomality constraint, $\mathbf{R}^T\mathbf{R} = \mathbf{I}$, is satisfied. It ensures the selection of only 1 loudspeaker for each direction. For the sake of simplicity, dropping the dependence on k and $\mathbf{\Theta}$, the optimization problem in (8) can be re-written and expressed as

$$\begin{aligned} \arg\min_{\mathbf{a},\mathbf{R}} \quad & \frac{1}{2}\sum_{q=1}^{Q} \|\mathbf{p}_q^t - \mathbf{X_q D a}\|_2^2 + \lambda\|\mathbf{R}\|_1 \\ \text{s.t.} \quad & \mathbf{D} = \mathbf{SR}, \quad \mathbf{R}^T\mathbf{R} = \mathbf{I}, \quad \mathbf{a}^T\mathbf{a} \leq \beta \end{aligned} \tag{10}$$

The above problem is non-convex in nature and can be solved by a splitting method based on Bregman iteration mentioned in [15]. The sparse iterative method proposed for this multi-zone scenario is discussed next.

3.2.1 Solution using Bregman Iteration

The steps of the iterative solution based on Bregman method [18] for the non-convex problem are given as

$$\begin{aligned} 1.\quad & (\mathbf{D}^\eta, \mathbf{R}^\eta) = \arg\min_{\mathbf{D},\mathbf{R}} \frac{1}{2}\sum_{q=1}^{Q} \|\mathbf{p}_q^t - \mathbf{X_q D a}^{\eta-1}\|_2^2 \\ & \qquad + \lambda\|\mathbf{R}\|_1 + \frac{\alpha}{2}\|\mathbf{R} - \mathbf{\Phi}^{\eta-1} + \mathbf{\Omega}^{\eta-1}\|_F^2 \\ & \qquad \text{s.t.} \qquad \mathbf{D} = \mathbf{SR} \\ 2.\quad & \mathbf{\Phi}^\eta = \arg\min_{\Phi} \frac{\alpha}{2}\|\mathbf{\Phi} - (\mathbf{R}^\eta + \mathbf{\Omega}^{\eta-1})\|_F^2 \\ & \qquad \text{s.t.} \quad \mathbf{\Phi^T\Phi} = \mathbf{I} \\ 3.\quad & \mathbf{\Omega}^\eta = \mathbf{\Omega}^{\eta-1} + \mathbf{R}^\eta - \mathbf{\Phi}^\eta \\ 4.\quad & \mathbf{a}^\eta \ = \mathbf{a}^*(\mathbf{D}^\eta) \end{aligned} \tag{11}$$

The sub-optimization problem in (11) is strictly convex with a closed form solution, as proposed in [15], given by

$$\mathbf{\Phi}_{opt}^\eta = \mathbf{U}\mathbf{V}\mathbf{\Sigma}^{-1/2}\mathbf{V}^T \tag{12}$$

where $\mathbf{U} = \mathbf{R}^\eta + \mathbf{\Omega}^{\eta-1}$ and $\mathbf{\Sigma}$ is a diagonal matrix obtained by the SVD factorization, $\mathbf{U}^T\mathbf{U} = \mathbf{V}\mathbf{\Sigma}\mathbf{V}^T$. Similar to [18], the additional linear constraint required to preserve the structure of first row of matrix $\mathbf{D}$ is given by

$$-\mathbf{1}^T\epsilon \leq (\mathbf{D}^T\mathbf{e} - \mathbf{1}^T Y_0^0) \leq \mathbf{1}^T\epsilon \tag{13}$$

Thus, the optimization problem in step-1 is re-framed as

$$\begin{aligned} (\mathbf{D}^\eta, \mathbf{R}^\eta) = \arg\min_{\mathbf{D},\mathbf{R}} \quad & \frac{1}{2}\sum_{q=1}^{Q} \|\mathbf{p}_q^t - \mathbf{X_q D a}^{\eta-1}\|_2^2 \\ & + \lambda\|\mathbf{R}\|_1 + \frac{\alpha}{2}\|\mathbf{R} - \mathbf{\Phi}^{\eta-1} + \mathbf{\Omega}^{\eta-1}\|_F^2 \\ \text{s.t.} \quad & \mathbf{D} = \mathbf{SR}, \quad -\mathbf{1}^T\epsilon \leq (\mathbf{D}^T\mathbf{e} - \mathbf{1}^T Y_0^0) \leq \mathbf{1}^T\epsilon \end{aligned} \tag{14}$$

Algorithm 1 Algorithm for reproducing the sound fields using the multi-zone sparse iterative method

1: Select the location O_q and radius r_q of each zone.
2: Obtain the target pressures p_q^t as defined in (6)
3: Create an overcomplete dictionary $\mathbf{S}$ for different $\mathbf{\Theta}$ by considering appropriate resolution of θ and ϕ
4: Initialize $\mathbf{a}^0$ and choose the appropriate values of parameter α, β, λ and ϵ
5: Initialize matrix $\mathbf{\Phi^0}$ with random values from normal distribution and set $\mathbf{\Omega}^0$ to a zero matrix.
6: **for** η = 1 to *ittr* **do**
7: Find $\mathbf{R}^\eta$ by solving the optimization problem (14)
8: Assign $\mathbf{U} = \mathbf{R}^\eta + \mathbf{\Omega}^{\eta-1}$
9: Compute the SVD form $\mathbf{U}^T\mathbf{U} = \mathbf{V}\mathbf{\Sigma}\mathbf{V}^T$
10: Update $\mathbf{\Phi}^\eta = \mathbf{U}\mathbf{V}\mathbf{\Sigma}^{-1/2}\mathbf{V}^T$
11: Update $\mathbf{\Omega}^\eta = \mathbf{\Omega}^{\eta-1} + \mathbf{R}^\eta - \mathbf{\Phi}^\eta$
12: Update $\mathbf{D}^\eta = \mathbf{SR}^\eta$
13: Solve problem defined in (8) and obtain $\mathbf{a}^*(\mathbf{D}^\eta)$
14: Update the loudspeaker weights $\mathbf{a}^\eta = \mathbf{a}^*(\mathbf{D}^\eta)$
15: **end for**
16: Obtain the loudspeaker location from the dictionary $\mathbf{S}$.
17: Generate the loudspeaker feeds using the obtained loudspeaker weights.

where, $\mathbf{e} = [1, 0, \cdots, 0]^T$ is a vector of size $(N_g+1)^2 \times 1$. Now, the loudspeaker feeds are computed iteratively and the final optimal weights are obtained when the convergence is achieved. The convergence here is the lowest required number of active loudspeakers. The steps involved in sparse iterative method are listed in Algorithm-1.

4. PERFORMANCE EVALUATION

The experimental setup considered to evaluate the performance of the proposed methodology is first described herein. Later, the analysis of the reconstructed sound fields on the basis of reproduction error along with the objective and subject evaluations are discussed.

4.1 Experimental Conditions

The problem of multi-zone reproduction is evaluated in a simulated environment considering $Q = 4$ spherical zones of reproduction each of radius $R_q = 0.1m$, $\forall q$. The corresponding local origins O_q are placed at the coordinates $(r_q, \Theta_q) = (0.875m, 90^o, \{\pm 149^o, \pm 31^o\})$, as shown in Fig. 1(a). A total of 512 loudspeakers are placed over a

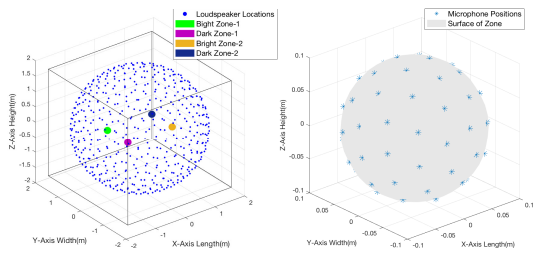


Figure 1: Figure illustrating (a) the experimental setup with loudspeaker arrangement and zone positions (b) considered microphone positions in each zone.

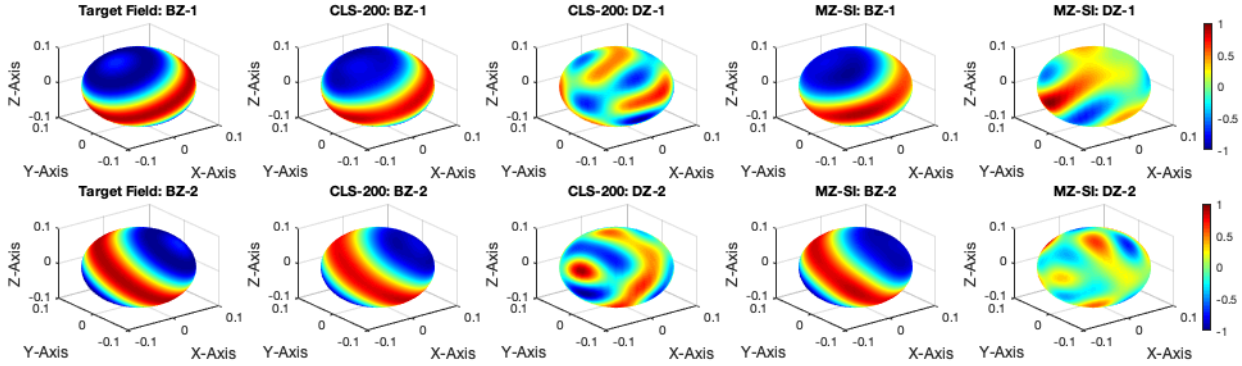


Figure 2: Figure illustrating the target sound fields in both bright zones along with the reproduced normalized sound fields using CLS and MZ-SI methods in different zones for $f = 2000$Hz.

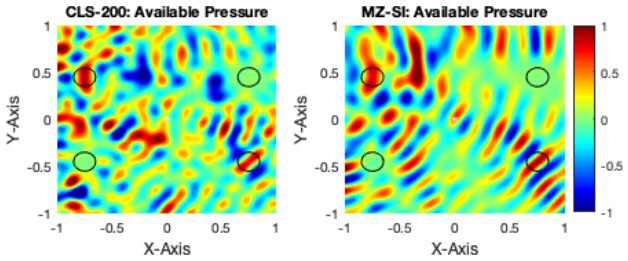


Figure 3: Figure illustrating the reproduced sound fields on a horizontal plane ($z = 0$) for $f = 2000$Hz.

sphere of radius $R_l = 2m$ following an icosahedron pattern. Similarily, 64 microphones are considered in each zone of reproduction as shown in Fig. 1(b). The personal zone problem is addressed for the scenario where 2 bright zones with two different angle of arrival of plane wave, $\Psi_1 = (150^o, 0^o)$ and $\Psi_2 = (45^o, -45^o)$, are considered. The proposed method is analysed for broadband frequency range from 500Hz$-$3.5KHz. In Algorithm-1, the sparsity parameter λ and the error parameter ϵ are set equal to 1 and 10^{-2}, respectively. Finally, to put an upper limit on the array effort, β is maintained equal to 1.

4.2 Analysis of Reconstructed Sound Fields

The sound fields in each bright zone are reconstructed by obtaining the loudspeaker feeds, $\mathbf{a}(k, \mathbf{\Theta}_L)$, corresponding to proposed multi-zone sparse iterative (MZ-SI) method and the constraint least-squares (CLS) problem defined in equation (8). For this multi-band scenario, MZ-SI methods significantly reduces the total active number of loudspeakers from 512 to 181. Therefore, for fare comparison, 200 loudspeakers rearranged in an icosahedron pattern are considered for CLS method. In Fig. 2, the reconstructed sound fields obtained using these methods over the surface of both the bright and the dark zones for $f = 2.0$KHz are illustrated. It can be seen that both the methods successfully reproduce the target sound fields in both bright zones. But MZ-SI method has advantage over CLS approach by reproducing more sparse sound fileds in both the dark zones. For clear visibility of acoustic contrast (AC), the reproduced sound field over a horizontal plane placed at $z = 0$m is presented in Fig. 3.

Under the considered methods, the obtained active loudspeaker positions and their corresponding weights are

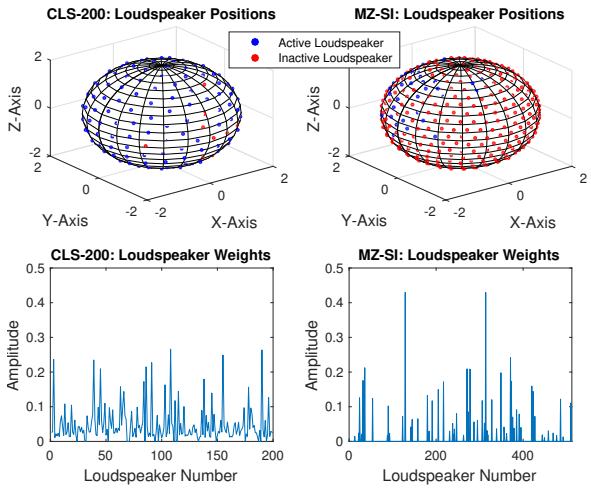


Figure 4: Figure illustrating the active and inactive loudspeaker positions along with their weights using CLS and MZ-SI methods for $f = 2000$Hz.

shown in Fig. 4. For $f = 2.0$KHz, the active number of loudspeakers using CLS and MZ-SI methods are 185 and 64, respectively. It is interesting to note that the number of active loudspeakers and their positions changes with the change in frequencies. Therefore, it is required to consider a common set of loudspeakers that minimizes the reproduction error across different frequency. Now, the optimal set of minimally distributed loudspeaker is given as

$$L^*_{opt} = \bigcup_{k \in \mathbb{K}} L_{active}(k) = \bigcup_{k \in \mathbb{K}} f(\mathbf{a}(k)) \qquad (15)$$

where, $f(\mathbf{a}(k)) = \{l \,|\, a_l(k) \neq 0,\ \forall l \in \{1 \dots L\}\}$ and $\mathbb{K}$ is a set of wave numbers k obtained for multiple f.

The accuracy of reproduction in each zone is determined by the root mean square error (RMSE) given by

$$RMSE_q(k) = \sqrt{\frac{1}{I_q} \sum_{\imath=1}^{I_q} |p^t_q(k, \mathbf{r}^\imath_\mathbf{q}) - p_q(k, \mathbf{r}^\imath_\mathbf{q})|^2} \qquad (16)$$

The overall RMSE is obtained by averaging individual RMSE across all zones. Similarly, the overall acoustic contrast is defined as an average of acoustic contrast obtained between each bright and each dark zone given by

$$AC(k) = \frac{I_B \mathbf{p}^T_B(k) \mathbf{p}_B(k)}{I_D \mathbf{p}^T_D(k) \mathbf{p}_D(k)} \qquad (17)$$

doi:10.25836/sasp.2019.13

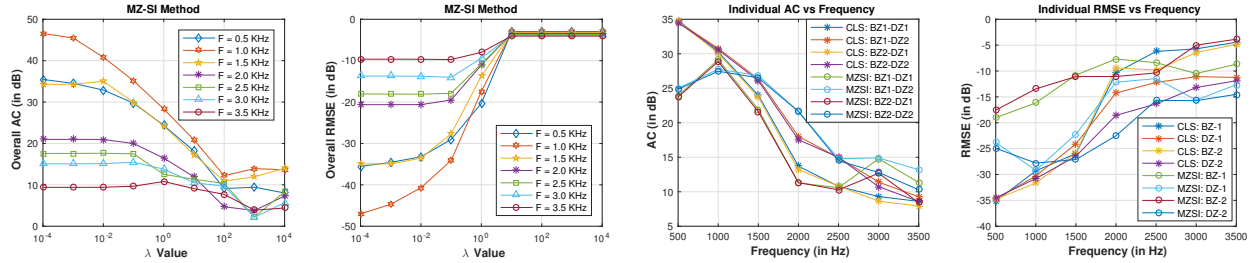


Figure 5: Figure illustrating the variation of (a) Overall AC, (b) Overall RMSE with respect to λ for MZ-SI method and the variation of (c) Individual AC, (d) Individual RMSE with respect to f for both CLS and MZ-SI methods with $\lambda = 1$.

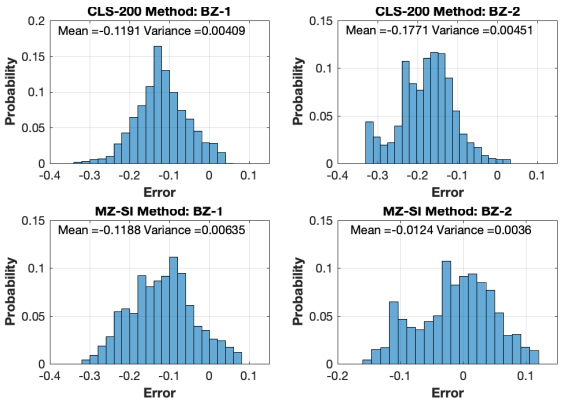


Figure 6: Figure illustrating the average error distribution obtained in both the bright zones for $f = 2000$Hz.

The variation of individual AC and individual RMSE for both the methods are shown in Fig. 5(c) and 5(d), respectively. From these figures, MZ-SI method shows improved acoustic contrasts in mid-rage frequencies and reduced error in dark zones when compared to CLS method. For higher frequencies, both methods shows similar performance. The improved performance of MZ-SI is the result of lower energy levels in DZ-2 and accurate reproduction in BZ-2. Fig. 6 supports the proposed approach in terms of average error distribution in both bright zones.

For the optimal choice of sparsity parameter λ, the variation of overall AC and overall RMSE is presented in Fig. 5(a) and 5(b). From the figures, it can be observed that for lower values of λ, the AC and RMSE shows good results while the performance degrades when too much of sparsity is introduced. Therefore, it is required to choose a particular λ which can withstand with the required AC and keep RMSE below a certain level.

4.3 Objective Evaluation

The reproduced multi-zone sound field is evaluated by measuring the scores obtained for PEAQ, PSM, PSMt and DI analysis mentioned in [19]- [20]. The reference signals are compared with the reconstructed signals in bright zones by using CLS methods and proposed MZ-SI approach. The obtained objective scores for the given test scenario are listed in Tab. 1. It can be observed that the obtained overall difference grades are similar for BZ-1 and better in BZ-2 when MZ-SI method is compared with CLS method. Hence, the reconstructed soundfield using MZ-SI can be considered perceptually better than soundfields obtained using CLS method.

Zone	Method	PEAQ	DI	PSM	PSMt
BZ-1	CLS	-3.441	-2.188	0.8763	0.5734
	MZ-SI	-3.493	-2.255	0.8694	0.5690
BZ-2	CLS	-3.578	-2.389	0.8367	0.5446
	MZ-SI	-3.363	-2.045	0.8830	0.5823

Table 1: Objective scores of reproduced multi-zone sound field using CLS and MZ-SI methods.

Zones	Methods	Naturalness		Presence		Preference		Source Envelope	
		μ	σ^2	μ	σ^2	μ	σ^2	μ	σ^2
BZ-1	CLS	3.1	0.82	2.2	1.57	2.6	1.33	2.4	0.99
	MZSI	3.0	0.83	2.3	0.98	3.1	1.09	2.4	0.90
BZ-2	CLS	2.9	0.78	2.3	1.21	2.9	0.93	2.4	1.10
	MZSI	3.3	1.38	2.7	1.21	3.0	1.10	2.6	0.86

Table 2: Obtained MOS scores for both the bright zones by evaluating 20 human subjects.

4.4 Subjective Evaluation

The proposed method is evaluated by obtaining of average mean opinion scores (MOS) provided by 20 human subject in consideration. Similar to [21], the subjects were asked to rate the quality of reconstructed signal on a scale of 1 to 5 for different spatial attributes [22], given as

- Naturalness: Audio listening true to real life
- Presence: Feel of presence in an environment
- Preference: Amount of pleasantness or harshness
- Source Envelope: Sound being around a person

The obtained MOS scores for both the methods under consideration are listed in Tab. 2. From the table, it can be observed that the objective scores supports the effectiveness of proposed MZ-SI method over CLS method.

5. CONCLUSION

In this paper, a framework for the multi-zone sound field reproduction via sparse expansion of spherical harmonics is presented. A sparse iterative method based on Bregman iteration to reproduce the target sound fields in multiple the bright zones is proposed herein. The proposed method provides a significant reduction in the active number of loudspeaker with improved acoustic contrast. The accuracy of the reproduced sound field is analysed by obtaining the average error distribution in the zones of interest. The objective and subjective evaluations also support the effectiveness of the proposed method. In future, the present work can be modified and extended to the reverberant environment. Also, an effort can be made to improve the computational complexity of the proposed sparse framework.

doi:10.25836/sasp.2019.13

6. ACKNOWLEDGEMENT

This work was funded by the SERB - DST under project no. SERB/EE/2017242

7. REFERENCES

[1] T. Betlehem, W. Zhang, M. A. Poletti, and T. D. Abhayapala, "Personal sound zones: Delivering interface-free audio to multiple listeners," *IEEE Signal Processing Magazine*, vol. 32, no. 2, pp. 81–91, 2015.

[2] J.-W. Choi and Y.-H. Kim, "Generation of an acoustically bright zone with an illuminated region using multiple sources," *The Journal of the Acoustical Society of America*, vol. 111, no. 4, pp. 1695–1700, 2002.

[3] J.-H. Chang, C.-H. Lee, J.-Y. Park, and Y.-H. Kim, "A realization of sound focused personal audio system using acoustic contrast control," *The Journal of the Acoustical Society of America*, vol. 125, no. 4, pp. 2091–2097, 2009.

[4] J. Cheer, S. J. Elliott, and M. F. S. Gálvez, "Design and implementation of a car cabin personal audio system," *Journal of the Audio Engineering Society*, vol. 61, no. 6, pp. 412–424, 2013.

[5] S. J. Elliott, J. Cheer, J.-W. Choi, and Y. Kim, "Robustness and regularization of personal audio systems," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 7, pp. 2123–2133, 2012.

[6] P. Coleman, P. J. Jackson, M. Olik, and J. Abildgaard Pedersen, "Personal audio with a planar bright zone," *The Journal of the Acoustical Society of America*, vol. 136, no. 4, pp. 1725–1735, 2014.

[7] M. Poletti, "An investigation of 2-d multizone surround sound systems," in *Audio Engineering Society Convention 125*, Audio Engineering Society, 2008.

[8] T. Betlehem and P. D. Teal, "A constrained optimization approach for multi-zone surround sound," in *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, pp. 437–440, IEEE, 2011.

[9] T. Betlehem and C. Withers, "Sound field reproduction with energy constraint on loudspeaker weights," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 8, pp. 2388–2392, 2012.

[10] W. Jin, W. B. Kleijn, and D. Virette, "Multizone soundfield reproduction using orthogonal basis expansion," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pp. 311–315, IEEE, 2013.

[11] Y. Cai, M. Wu, and J. Yang, "Sound reproduction in personal audio systems using the least-squares approach with acoustic contrast control constraint," *The Journal of the Acoustical Society of America*, vol. 135, no. 2, pp. 734–741, 2014.

[12] Y. J. Wu and T. D. Abhayapala, "Multizone 2d soundfield reproduction via spatial band stop filters," in *Applications of Signal Processing to Audio and Acoustics, 2009. WASPAA'09. IEEE Workshop on*, pp. 309–312, IEEE, 2009.

[13] Y. J. Wu and T. D. Abhayapala, "Spatial multizone soundfield reproduction: Theory and design," *IEEE Transactions on audio, speech, and language processing*, vol. 19, no. 6, pp. 1711–1720, 2011.

[14] W. Zhang, T. D. Abhayapala, T. Betlehem, and F. M. Fazi, "Analysis and control of multi-zone sound field reproduction using modal-domain approach," *The Journal of the Acoustical Society of America*, vol. 140, no. 3, pp. 2134–2144, 2016.

[15] R. Lai and S. Osher, "A splitting method for orthogonality constrained problems," *Journal of Scientific Computing*, vol. 58, no. 2, pp. 431–449, 2014.

[16] D. B. Ward and T. D. Abhayapala, "Reproduction of a plane-wave sound field using an array of loudspeakers," *IEEE Transactions on speech and audio processing*, vol. 9, no. 6, pp. 697–707, 2001.

[17] B. Rafaely, *Fundamentals of spherical array processing*, vol. 8. Springer, 2015.

[18] S. N. Kalkur, S. Reddy C, and R. M. Hegde, "Joint source localization and separation in spherical harmonic domain using a sparsity based method," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[19] T. Thiede, W. C. Treurniet, R. Bitto, C. Schmidmer, T. Sporer, J. G. Beerends, and C. Colomes, "Peaq-the itu standard for objective measurement of perceived audio quality," *Journal of the Audio Engineering Society*, vol. 48, no. 1/2, pp. 3–29, 2000.

[20] R. Huber and B. Kollmeier, "Pemo-qa new method for objective audio quality assessment using a model of auditory perception," *IEEE Transactions on audio, speech, and language processing*, vol. 14, no. 6, pp. 1902–1911, 2006.

[21] A. Dagar, S. S. Nitish, and R. Hegde, "Joint adaptive impulse response estimation and inverse filtering for enhancing in-car audio," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 416–420, April 2018.

[22] C. S. Reddy and R. M. Hegde, "Horizontal plane hrtf interpolation using linear phase constraint for rendering spatial audio," in *Signal Processing Conference (EUSIPCO), 2016 24th European*, pp. 1668–1672, IEEE, 2016.

doi:10.25836/sasp.2019.13

PARAMETRIC FIRST-ORDER AMBISONIC DECODING FOR HEADPHONES UTILISING THE CROSS-PATTERN COHERENCE ALGORITHM

Leo McCormack and Symeon Delikaris-Manias
Aalto University, Department of Signal Processing and Acoustics, Espoo, Finland
leo.mccormack@aalto.fi

ABSTRACT

Binaural ambisonics decoding is a means of reproducing a captured or synthesised sound-field, as described by a spherical harmonic representation, over headphones. The majority of ambisonic decoders proposed to date are based on a signal-independent approach; operating via a linear mapping between the input spherical harmonic signals and the output binaural signals. While this approach is computationally efficient, an impractically high input order is often required to deliver a sufficiently accurate rendition of the original spatial cues to the listener. This is especially problematic, as the vast majority of commercially available Ambisonics microphones are first-order, which ultimately results in numerous perceptual deficiencies during reproduction. Therefore, in this paper, a signal-dependent and parametric binaural ambisonic decoder is proposed, which is specifically intended to reproduce first-order input with high perceptual accuracy. The proposed method assumes a sound-field model of one source and one non-isotropic ambient component per narrow-band. It then employs the Cross-Pattern Coherence (CroPaC) post-filter, in order to segregate these components with improved spatial selectivity. Listening test results indicate that the proposed method, when using first-order input, performs similarly to third-order Ambisonics reproduction.

1. INTRODUCTION

Reproduction of synthesised or captured sound-fields is an important component in many immersive audio applications, where flexibility, in terms of both content generation and playback setup, is highly favoured. Methods formulated in the spherical harmonic domain (SHD) [1] are often well-suited to this task, as the recording and reproduction operations may be decoupled; with spherical harmonic signals serving as an intermediary. This SHD-based ecosystem for sound-field capture and reproduction is popularly known as Ambisonics [2], where the generation of spherical harmonic signals and the subsequent reproduction of the sound scene that they describe, is referred to as ambisonic encoding and decoding, respectively. Regarding the latter, currently proposed decoders may be loosely categorised as either non-parametric (signal-independent) or parametric (signal-dependent). Non-parametric binaural reproduction relies on a complex, frequency-dependent, and linear mapping of the input signals to the binaural channels. Whereas parametric methods operate by imposing a set of assumptions regarding the composition of the sound-field and are signal-dependent. Methods that fall within this latter category often rely on the extraction of perceptually meaningful parameters in the time-frequency domain, with the aim of mapping input signals to the binaural channels in a more informed manner [3]. This paper is primarily concerned with parametric reproduction of first-order spherical harmonic input over headphones. A binaural decoder is proposed for this task, which utilises the Cross-Pattern Coherence (CroPaC) [4] spatial post-filter; in order to segregate the sound-field into one source and one non-isotropic ambient component per time-frequency tile during the analysis stage. The method then employs the optimal mixing approach described in [5], to synthesise the output binaural signals.

1.1 Non-parametric binaural ambisonics decoding

Binaural ambisonics decoding is conducted via the application of a matrix of filters, which appropriately maps the input spherical harmonic signals to the binaural channels in a linear manner. Therefore, no time-varying distortions are introduced into the output signals. The decoding filters may be derived by approximating the directivity patterns of the listener's head-related transfer functions (HRTF), using the spherical harmonic basis functions, in a least-squares (LS) sense. However, in order to sufficiently approximate and reproduce these complicated directional patterns, a dense grid of HRTF measurements and a high input order is required; often in the range of 15-20th order. For practical reasons, the input order is typically truncated to a much lower order than that of the spatial order of the HRTF measurement grid. This, in turn, results in direction-dependent timbral colourations in the binaural signals. In addition, Ambisonics reproduction is inherently limited by the spatial resolution of the input format. For lower-orders, this has been found to exhibit numerous perceptual deficiencies, including: localisation ambiguity, comb-filtering effects, poor externalisation, and a loss of envelopment [6–10].

Timbral colourations due to input order truncation es-

© Leo McCormack and Symeon Delikaris-Manias. Licensed under a Creative Commons Attribution 4.0 International License (CC BY 4.0). **Attribution:** Leo McCormack and Symeon Delikaris-Manias. "Parametric first-order ambisonic decoding for headphones utilising the Cross-Pattern Coherence algorithm", 1st EAA Spatial Audio Signal Processing Symposium, Paris, France, 2019.

doi:10.25836/sasp.2019.26

pecially affect high-frequencies, since the high-frequency energy is predominantly concentrated in the higher-order components. This loss of energy may be compensated for via diffuse-field equalisation [11]. However, this loss of high-frequency energy is largely due to the miss-match between the input order and the spatial order of the measurement grid, which is directly proportional to its density. Therefore, rather than applying post-equalisation filters, one may simply reduce the number of points in the HRTF measurement grid, such that its spatial order is more in-line with that of the input order; as suggested in [12]. This approach is often referred to as Spatial Re-sampling (SPR) or virtual loudspeaker decoding. In this case, rather than assigning high-frequency energy to higher-order components and subsequently discarding it, due to order truncation, the energy is instead aliased back into the lower-order components and preserved. However, while this approach improves upon the perceived timbral short-comings of lower-order binaural Ambisonic reproduction, it does not eliminate them, nor does it address the spatial deficiencies of the method.

The localisation ambiguities associated with lower-order binaural Ambisonic reproduction are due to a degradation of the reproduced binaural cues. There are two key causes for this. The first is due to the inherent low input spatial resolution, which leads to erroneously high signal coherence between the output channels; when generated in a linear manner. Whereas the second is due to the LS decoder itself, as it is unable to sufficiently fit the lower-order spherical harmonic patterns to the highly directive HRTF patterns. To address this latter limitation, an alternative method was proposed in [13], which conducts preliminary time-alignment of the Head-related impulse responses (HRIRs) and performs the LS fitting with an additional diffuse-field coherence constraint. The method essentially exploits prior knowledge of the bandwidth in which the inter-aural level differences (ILDs) are the dominant localisation cues; which is above approximately 1.5 kHz, as described by the Haas effect [14]. By discarding the phase information of the HRTFs at frequencies above 1.5 kHz, the LS fitting instead prioritises the delivery of the correct magnitude responses; rather than the phase. Thus it ultimately yields improved ILD cues and diminished inter-aural time difference (ITD) cues; but in a frequency range where ILD cues are more important for localisation. The same principle was also later employed in [15]. However, while these aforementioned approaches yield considerable improvements over traditional decoders, as shown with formal listening tests [13, 16], their performance with first-order input still deviates from that of higher-orders and directly binauralised scenes. This is especially problematic as the vast majority of commercially available Ambisonic microphones and available content are first-order.

1.2 Parametric binaural decoding

The inherent perceptual limitations associated with lower-order Ambisonics are primarily as a result of the erroneously high coherence between the output channels. In order to overcome these limitations, signal-dependent and parametric alternatives have been proposed [17–22]. These methods employ a sound-field model, which lays out a set of assumptions regarding the composition of the sound-field. The methods operate by extracting perceptually meaningful parameters in the time-frequency domain, and often employ dedicated rendering techniques for different components. The two main challenges when designing a parametric method are therefore: 1) identifying a perceptually robust sound-field model, and 2) employing the appropriate signal processing techniques in order to realise the model, with minimal artefacts incurred.

The most well-known and established parametric reproduction method is Directional Audio Coding (DirAC) [17], which employs a sound-field model consisting of one plane-wave and one diffuseness estimate per time-frequency tile. These parameters are derived from the active-intensity vector, in the case of first-order input. The plane-wave components are panned directly to the loudspeakers using Vector-base Amplitude Panning (VBAP) [23], and the diffuse components are sent to all loudspeakers and decorrelated. More recent formulations of DirAC also allow for multiple plane-wave and diffuseness estimates via spatially-localised active-intensity vectors, using higher-order input [18, 19]. In [24], a post-filter was proposed, which adaptively mixes between linearly decoded output and DirAC rendered outputs, in order to improve the output signal fidelity.

High Angular Resolution plane-wave Expansion (HARPEX) [20], is another example of a parametric method, which operates by extracting two plane-wave components per frequency using first-order input. The Sparse-Recovery method [21] extracts a number of plane-waves, which corresponds to up to half the number of input channels of arbitrary order. The COding and Multi-Parameterisation of Ambisonic Sound Scenes (COMPASS) method [25] also extracts multiple source components; up to half the number of input channels. However, it employs an additional residual stream that encapsulates the remaining diffuse and ambient components in the scene. An alternative parameterisation of the sound-field was also presented in [26, 27], which circumvents the modelling of the sound-field with conventional parameters, such as source direction or diffuseness. It considers only the perceived quality of the individual output channels and the perceived quality of the spatial attributes of the reproduction. In [22] an adaptive LS-binaural decoder was proposed, which constrains the covariance matrix of the decoded audio to reflect that of the covariance matrix of the listeners HRTFs. The method does not require explicit estimation of the direct/diffuse balance, nor does it need to employ de-correlation.

1.2.1 Parametric binaural decoding with optimal mixing

The main point of criticism regarding parametric reproduction methods is the incursion of time-varying artefacts. These occur either due to the input scene not conforming to the assumed sound-field model, or due to the rendering techniques not being sufficiently robust. Therefore, in [28], an *optimal mixing* technique was proposed, which attempts to synthesise the output using a linear combination of linearly decoded prototype signals, as much as possible; thus

 doi:10.25836/sasp.2019.26

retaining much of the high single-channel fidelity in the output. The method then employs decorrelation only if the output inter-channel dependencies still deviate from the target, thus mitigating potential decorrelation artefacts; such as the temporal smearing of transients.

The approach is formulated in the covariance domain and relies on the construction of time-varying narrowband target covariance matrices, which are dictated by the sound-field model of the parametric method. It then employs traditional linearly decoded audio as a prototype. The approach has been employed previously in [19] using the DirAC sound-field model, and also in the closed-form solution of [22] for binaural reproduction. However, in principle, any sound-field model may employ this optimal mixing approach for the synthesis stage.

1.3 Motivation for this work

Further development of first-order ambisonic decoders is of particular interest; given the prevalence of first-order commercially available Ambisonic microphone arrays and material accessible on the internet. Recent advancements have shown improvements in the perceived performance of linear binaural ambisonic decoding [13, 15]. However, these decoders still rely on input orders which may be considered impractical for wide-spread adoption today. Especially given the quadratic growth in the number of channels (and microphones) with increasing order, leading to more expensive arrays and higher bandwidths. Therefore, the need for robust signal-dependent parametric alternatives for lower-orders is well established.

In this paper, a first-order binaural decoder is proposed [1], which employs a parametric sound-field model that assumes one source and one directional ambient component per time-frequency tile. The model is inspired by the multi-source and non-isotropic ambient model employed by the COMPASS method [25]. However, given the low-resolution of first-order input, the proposed method employs an additional CroPaC spatial post-filter [4] to better isolate the source component, and an inverse CroPaC post-filter to obtain the ambient component. The proposed method also extracts instantaneous direction-of-arrival (DoA) estimates via steered-response power (SRP) activity-maps and peak-finding; thus forgoing the need for long temporal averaging of input covariance matrices, as required by sub-space alternatives. The method also employs the optimal mixing solution of [5], in order to minimise the amount of decorrelated signal energy in the output, and synthesise the target inter-channel dependencies in a linear manner, as much as possible.

1.4 Organisation of the paper

Section 2 describes linear binaural ambisonic decoding, which is employed as a basis for the proposed method. Section 3 details the proposed method. Subjective evaluation of the proposed algorithm, through listening tests, is described in Section 4, and Section 5 concludes the paper.

[1] A VST audio plug-in of the proposed decoder may be found here: http://research.spa.aalto.fi/publications/papers/sasp19-parametric/

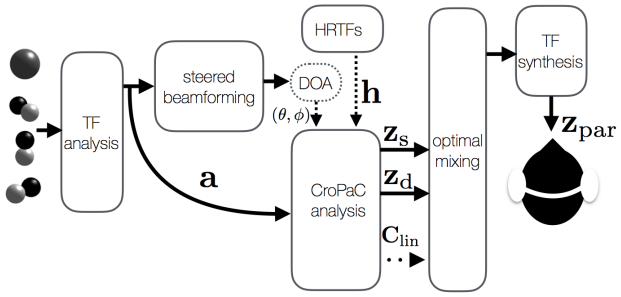


Figure 1: Overall block diagram of the proposed method.

2. LINEAR BINAURAL AMBISONICS DECODING

Due to the signal- and frequency-dependent nature of the proposed algorithm (described in Section 3) it is assumed that the input ambisonic signals have been transformed into the time-frequency (t, f) domain; where t and f denote the time and frequency indices, respectively. The ambisonic signals, $\mathbf{a}$, may be synthesised by mapping monophonic signals onto spherical harmonic basis functions or captured utilising a microphone array with subsequent suitable encoding

$$\mathbf{a} = [a_{00}, a_{1(-1)}, a_{10}, \ldots, a_{N(N-1)}, a_{NN}]^{\mathrm{T}} \in \mathbb{C}^{(N+1)^2 \times 1}, \tag{1}$$

where a_{nm} are the individual ambisonic signals for each order, n, and degree, m, up to the maximum order, N. It is assumed, henceforth, that the input ambisonic signals conform to the ortho-normalised (N3D) and Ambisonic Channel Numbering (ACN) conventions.

Ambisonic signals may be decoded for headphone playback as

$$\mathbf{z}_{\mathrm{lin}}(t, f) = \mathbf{D}_{\mathrm{bin}}(f)\mathbf{a}(t, f), \tag{2}$$

where $\mathbf{z}_{\mathrm{lin}}(t, f) \in \mathbb{C}^{2\times 1}$ are the output binaural signals, and $\mathbf{D}_{\mathrm{bin}}(f) \in \mathbb{C}^{2\times(N+1)^2}$ is a binaural ambisonic decoding matrix, derived using one of a number of approaches [12, 13, 15].

3. PROPOSED PARAMETRIC BINAURAL AMBISONICS DECODER

The proposed first-order decoder employs a sound-field model comprising one source component and one non-isotropic ambient component per time-frequency tile. The method first estimates the source DoA via steered-response power (SRP) beamforming and subsequent peak-finding. The source stream is then segregated by steering a beamformer toward the estimated DoA, and employing an additional CroPaC post-filtering operation to improve its spatial selectivity. The ambient stream is then simply the residual, once the source component has been subtracted from the input sound-field. The two streams are then binauralised and fed into an optimal mixing unit, along with the ambisonic prototype covariance matrix, in order to generate the binaural output. A block diagram of the proposed method is depicted in Fig. 1.

3.1 Analysis

3.1.1 DoA estimation

A DoA estimator based on SRP beamforming and peak-finding [29] is suggested for the proposed rendering method. Note that the chosen scanning grid should preferably take the angular resolution of human hearing into account [14]. This approach yields instantaneous estimates of the source direction, (θ_s, ϕ_s), which is in keeping with the instantaneous CroPaC post-filter values. Therefore, the need for temporal averaging of input covariance matrices is avoided in the analysis stage, as would be required by subspace-based alternatives.

3.1.2 CroPaC post-filter

The cross-correlation between the omni and dipole may be utilised as a spatial post-filter [4]

$$G(t,f) = \frac{2}{\sqrt{3}}\Re[{a_{00}}^*(t,f)a_{11}(t,f)], \tag{3}$$

where $\Re$ denotes the real operator and * denotes the complex conjugate. This value is then normalised with the input sound-field energy and half-wave rectified

$$\hat{G}(t,f) = \max\left[\frac{G(t,f)}{|a_{00}(t,f)|^2 + \sum_{-1}^{1}|a_{1m}(t,f)/\sqrt{3}|^2}, \lambda\right], \tag{4}$$

where $\lambda \in [0, ..., 1]$ is a parameter which influences the severity of the post-filter, similarly to the spectral floor of a traditional post-filter. Note that the spatial selectivity of the CroPaC post-filter is sharper than a conventional first-order beam pattern, due to the multiplication (rather than summation) of the two signals in (3).

The CroPaC spatial filter may be steered in the source direction, by first rotating the spherical harmonic signals using an appropriate rotation matrix, $\mathbf{M}_{\text{rot}}(\theta_s, \phi_s) \in \mathbb{R}^{(N+1)^2\times(N+1)^2}$, as

$$\mathbf{a}_{\text{rot}}(t,f) = \mathbf{M}_{\text{rot}}(\theta_s, \phi_s)\mathbf{a}(t,f), \tag{5}$$

where $\mathbf{a}_{\text{rot}}(t,f) \in \mathbb{C}^{(N+1)^2\times 1}$ are the resulting rotated signals, which are then employed for the post-filter estimation (3). For details regarding the calculation of this rotation matrix, the reader is directed to [30].

3.2 Synthesis

The source stream, $\mathbf{z}_s(t,f)$, is obtained by directly applying the HRTF gains to the extracted source signal as

$$\mathbf{z}_s(t,f) = \mathbf{h}\frac{\hat{G}(t,f)}{(N+1)^2}\mathbf{w}^{\text{T}}\mathbf{a}(t,f), \tag{6}$$

where $\mathbf{h} \in \mathbb{C}^{2\times 1}$ and $\mathbf{w} \in \mathbb{R}^{(N+1)^2\times 1}$ are HRTF gains and static beamforming weights [1], corresponding to the analysed DoA at each time-frequency tile, respectively. A visual depiction of the improved spatial selectivity of the source beamformer with the post-filter is given in Fig. 2.

The ambient stream, $\mathbf{z}_d(t,f)$, is then obtained by subtracting the source signal from the input scene, and decoding it to headphones using a binaural ambisonic decoder

$$\mathbf{z}_d(t,f) = \mathbf{D}_{\text{bin}}[\mathbf{a}(t,f) - \frac{\hat{G}(t,f)}{(N+1)^2}\mathbf{y}\mathbf{w}^{\text{T}}\mathbf{a}(t,f)], \tag{7}$$

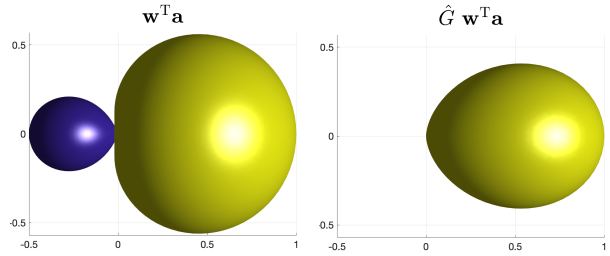


Figure 2: A first-order hyper cardioid beamformer without (left) and with (right) the CroPaC post-filter ($\lambda = 0$).

where $\mathbf{y} \in \mathbb{R}^{(N+1)^2\times 1}$ are the spherical harmonic weights for the analysed DoA.

Optionally, the ambient stream may be decorrelated, in order to further minimise the inter-channel coherence between the binaural channels

$$\hat{\mathbf{z}}_d(t,f) = \mathcal{D}\left[\mathbf{z}_d(t,f)\right] \tag{8}$$

where $\mathcal{D}[.]$ denotes a decorrelation operation on the enclosed signals.

3.2.1 Optimal mixing

Rather than summing the source (6) and ambient (8) streams together, to acquire the binaural output, an alternative synthesis approach is suggested. This approach is based on the covariance domain framework, termed here as *optimal mixing*, and is described in [5, 28]. The method employs linearly decoded signals, $\mathbf{z}_{\text{lin}}(t,f)$, as a prototype, which has the following base-line covariance matrix (note that the time-frequency indices have been omitted for the brevity of notation)

$$\mathbf{C}_{\text{lin}} = \text{E}[\mathbf{z}_{\text{lin}}\mathbf{z}_{\text{lin}}^{\text{H}}] = \mathbf{D}_{\text{bin}}\mathbf{C}_{\text{a}}\mathbf{D}_{\text{bin}}^{\text{H}}, \tag{9}$$

where $\mathbf{C}_{\text{a}} = \text{E}[\mathbf{a}\mathbf{a}^{\text{H}}] \in \mathbb{C}^{(N+1)^2\times(N+1)^2}$ is the covariance matrix of the input signals.

A time-varying and narrow-band target covariance matrix is then required, which may be defined in this case as

$$\mathbf{C}_{\text{target}} = \text{E}[(\mathbf{z}_s + \mathbf{z}_d)(\mathbf{z}_s + \mathbf{z}_d)^{\text{H}}]. \tag{10}$$

The optimal mixing solution then provides the values for matrices $\mathbf{A} \in \mathbb{C}^{2\times 2}$ and $\mathbf{B} \in \mathbb{C}^{2\times 2}$, in the following equation

$$\mathbf{A}\mathbf{C}_{\text{lin}}\mathbf{A}^{\text{H}} + \mathbf{B}\tilde{\mathbf{C}}_{\text{lin}}\mathbf{B}^{\text{H}} = \mathbf{C}_{\text{target}} \tag{11}$$

where $\tilde{\mathbf{C}}_{\text{lin}} = \text{diag}\left[\mathcal{D}\left[\mathbf{z}_{\text{lin}}\right]\mathcal{D}\left[\mathbf{z}_{\text{lin}}\right]^{\text{H}}\right]$ is a diagonal matrix consisting of the diagonal entries of the covariance matrix of a decorrelated version of the linearly decoded prototype. More information regarding the derivation and calculation of these mixing matrices is given in [5].

The output audio $\mathbf{z}_{\text{par}}$ may then be obtained as

$$\mathbf{z}_{\text{par}} = \mathbf{A}\mathbf{z}_{\text{lin}} + \mathbf{B}\mathcal{D}\left[\mathbf{z}_{\text{lin}}\right], \tag{12}$$

which ideally exhibits all of the target inter-channel dependencies, as dictated by the target covariance matrix $\text{E}[\mathbf{z}_{\text{par}}\mathbf{z}_{\text{par}}^{\text{H}}] \simeq \mathbf{C}_{\text{target}}$. However, note that in practice, due to the need for regularisation of the input covariance matrices, the solution is never exact.

doi:10.25836/sasp.2019.26

4. EVALUATION

A multiple-stimulus test was conducted, in order to compare the perceived performance of the proposed approach (*CroPaC1*), with both first-order (*Ambi1*) and third-order (*Ambi3*) spatial re-sampling ambisonic decoders [12]. Note that the optimal mixing approach (12) was employed, using the same ambisonic decoder for the prototype $\mathbf{z}_{\mathrm{lin}}$.

Synthetic test scenes were created as follows: multiple speakers (*speakers*), a modern funk band (*groove*), and a mix comprising of a speaker, clapping, water fountain and piano (*mix*). The individual sources were binauralised directly for the reference test cases, and encoded into first- and third- order signals and passed through their respective decoders for the *_dry* test cases. A shoebox image-source room simulator [2] was employed to obtain reverberant (*_rev*) counterparts. The room simulator was configured to resemble a small hall. The image-sources arriving at the receiver position were binauralised directly for the reference cases, and encoded into spherical harmonic signals for the reverberant test cases.

The evaluation consisted of three parts, which addressed the spatial, timbral, and overall differences between the methods. For the evaluation of **spatial** differences, the mean spectra of the reference was imposed onto the test cases. Therefore, timbral differences between the methods were greatly mitigated, but still retained their original spatial characteristics. The test subjects were explicitly instructed to ignore any remaining timbral differences, which may not have been addressed by the equalisation. The anchor was obtained as an omni-directional spherical harmonic component, replicated to each binaural channel and also equalised. For evaluating the **timbral** differences, the mean spectra of each test case, was imposed onto an omni-directional signal and sent to both left and right channels. Therefore, the spatial differences between the methods were eliminated, thus retaining only the timbral differences. The anchor was obtained as the mean spectra of the reference, replicated to each binaural channel and low-passed filtered at 4 kHz. Finally, for assessing the **overall** differences, only the broad-band RMS of the reference was used to normalise the test cases. Therefore, all timbral and spatial differences between test cases remained, and the test participants graded the samples based on their subjective weighting of the reproduced attributes.

A total of 14 expert listeners participated in the listening tests in purpose-built headphone booths. The present authors did not take part in the tests. The means and 95% confidence intervals of the results are shown in Fig. 3. It can be observed that the spatial and timbral characteristics of the reference were more closely reproduced using the proposed method, when compared to conventional ambisonics with the same first-order input. Furthermore, the proposed method yielded scores more inline with that of third-order ambisonics. It should be highlighted that third-order ambisonics employs four times the number of input channels than that of the proposed method. This, therefore, represents a significant reduction in bandwidth, without compromising the perceived spatial accuracy or fidelity.

[2] The shoe-box room simulator employed for the reverberant test cases may be found here: https://github.com/polarch/shoebox-roomsim

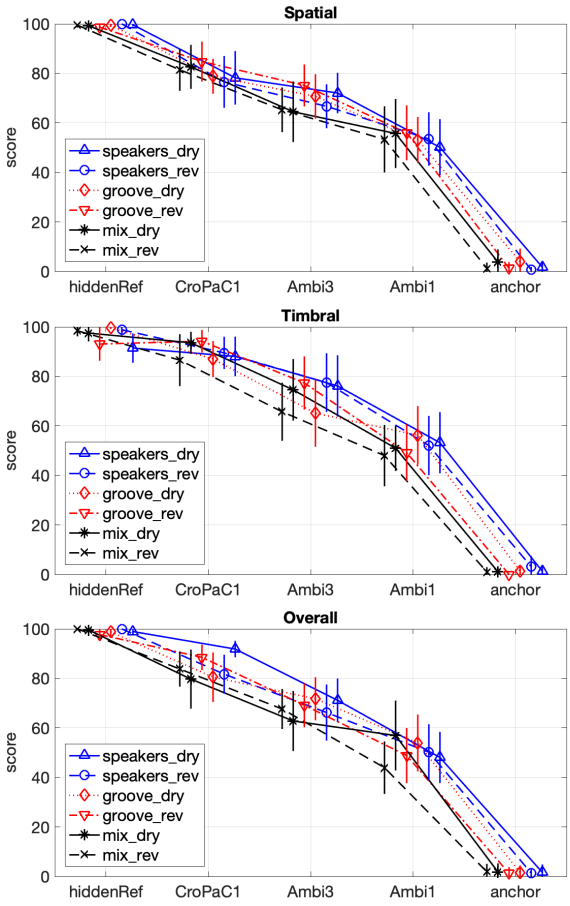


Figure 3: The means and 95% confidence intervals for each individual sound scene. The evaluation criteria were: spatial only, timbral only, and overall (top-bottom).

5. CONCLUSION

This paper has proposed a first-order parametric binaural ambisonic decoder, which employs a sound-field model comprising one source and one directional ambient component per time-frequency tile. The proposed approach first isolates the source components using a spherical harmonic domain beamformer, modulated by the Cross-Pattern Coherence (CroPaC) spatial post-filter. The ambient stream is then simply the residual, once the source components have been subtracted from the input sound-field. The proposed approach is inspired by the COMPASS method [25]. However, along with the CroPaC post-filter, it also employs instantaneous source direction estimation and synthesises the output in a linear manner as much as possible; in order to improve the fidelity of the output signals. Formal listening tests indicate that the proposed first-order decoder performs similarly to (or exceeds) third-order spatial re-sampling ambisonics decoding, in terms of the perceived spatial and timbral attributes of the reproduction.

6. ACKNOWLEDGEMENTS

This research has received funding from the Aalto ELEC Doctoral School. Many thanks are also extended to Dr. Archontis Politis for the insightful discussions regarding the method.

7. REFERENCES

[1] B. Rafaely, *Fundamentals of spherical array processing*, vol. 8. Springer, 2015.

[2] M. A. Gerzon, "Periphony: With-height sound reproduction," *Journal of the Audio Engineering Society*, vol. 21, no. 1, pp. 2–10, 1973.

[3] V. Pulkki, S. Delikaris-Manias, and A. Politis, "Parametric time-frequency domain spatial audio," 2018.

[4] S. Delikaris-Manias and V. Pulkki, "Cross pattern coherence algorithm for spatial filtering applications utilizing microphone arrays," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 11, pp. 2356–2367, 2013.

[5] J. Vilkamo and T. Backstrom, "Time–frequency processing: Methods and tools," in *Parametric Time-Frequency Domain Spatial Audio*, pp. 3–23, John Wiley & Sons, 2018.

[6] O. Santala, H. Vertanen, J. Pekonen, J. Oksanen, and V. Pulkki, "Effect of listening room on audio quality in ambisonics reproduction," in *Audio Engineering Society Convention 126*, Audio Engineering Society, 2009.

[7] S. Braun and M. Frank, "Localization of 3d ambisonic recordings and ambisonic virtual sources," in *1st International Conference on Spatial Audio,(Detmold)*, 2011.

[8] A. Avni, J. Ahrens, M. Geier, S. Spors, H. Wierstorf, and B. Rafaely, "Spatial perception of sound fields recorded by spherical microphone arrays with varying spatial resolution," *The Journal of the Acoustical Society of America*, vol. 133, no. 5, pp. 2711–2721, 2013.

[9] S. Bertet, J. Daniel, E. Parizet, and O. Warusfel, "Investigation on localisation accuracy for first and higher order ambisonics reproduced sound sources," *Acta Acustica united with Acustica*, vol. 99, no. 4, pp. 642–657, 2013.

[10] P. Stitt, S. Bertet, and M. van Walstijn, "Off-centre localisation performance of ambisonics and hoa for large and small loudspeaker array radii," *Acta Acustica united with Acustica*, vol. 100, no. 5, pp. 937–944, 2014.

[11] Z. Ben-Hur, F. Brinkmann, J. Sheaffer, S. Weinzierl, and B. Rafaely, "Spectral equalization in binaural signals represented by order-truncated spherical harmonics," *The Journal of the Acoustical Society of America*, vol. 141, no. 6, pp. 4087–4096, 2017.

[12] B. Bernschütz, A. V. Giner, C. Pörschmann, and J. Arend, "Binaural reproduction of plane waves with reduced modal order," *Acta Acustica united with Acustica*, vol. 100, no. 5, pp. 972–983, 2014.

[13] M. Zaunschirm, C. Schörkhuber, and R. Höldrich, "Binaural rendering of ambisonic signals by head-related impulse response time alignment and a diffuseness constraint," *The Journal of the Acoustical Society of America*, vol. 143, no. 6, pp. 3616–3627, 2018.

[14] J. Blauert, *Spatial hearing: the psychophysics of human sound localization*. MIT press, 1997.

[15] C. Schörkhuber, M. Zaunschirm, and R. Höldrich, "Binaural rendering of ambisonic signals via magnitude least squares," in *Proceedings of the DAGA*, vol. 44, pp. 339–342, 2018.

[16] H. Lee, M. Frank, and F. Zotter, "Spatial and timbral fidelities of binaural ambisonics decoders for main microphone array recordings," in *Audio Engineering Society Conference: 2019 AES International Conference on Immersive and Interactive Audio*, Audio Engineering Society, 2019.

[17] V. Pulkki, "Directional audio coding in spatial sound reproduction and stereo upmixing," in *Audio Engineering Society Conference: 28th International Conference: The Future of Audio Technology–Surround and Beyond*, Audio Engineering Society, 2006.

[18] A. Politis, J. Vilkamo, and V. Pulkki, "Sector-based parametric sound field reproduction in the spherical harmonic domain," *IEEE Journal of Selected Topics in Signal Processing*, vol. 9, no. 5, pp. 852–866, 2015.

[19] A. Politis, L. McCormack, and V. Pulkki, "Enhancement of ambisonic binaural reproduction using directional audio coding with optimal adaptive mixing," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics,*, 2017.

[20] S. Berge and N. Barrett, "High angular resolution planewave expansion," in *Proc. of the 2nd International Symposium on Ambisonics and Spherical Acoustics May*, pp. 6–7, 2010.

[21] A. Wabnitz, N. Epain, and C. T. Jin, "A frequency-domain algorithm to upscale ambisonic sound scenes," in *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 385–388, IEEE, 2012.

[22] C. Schörkhuber and R. Höldrich, "Linearly and quadratically constrained least-squares decoder for signal-dependent binaural rendering of ambisonic signals," in *Audio Engineering Society Conference: 2019 AES International Conference on Immersive and Interactive Audio*, Audio Engineering Society, 2019.

[23] V. Pulkki, "Virtual sound source positioning using vector base amplitude panning," *Journal of the audio engineering society*, vol. 45, no. 6, pp. 456–466, 1997.

[24] O. Thiergart, G. Milano, and E. A. Habets, "Combining linear spatial filtering and non-linear parametric processing for high-quality spatial sound capturing," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 571–575, IEEE, 2019.

[25] A. Politis, S. Tervo, and V. Pulkki, "Compass: Coding and multidirectional parameterization of ambisonic sound scenes," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6802–6806, IEEE, 2018.

[26] J. Vilkamo and S. Delikaris-Manias, "Perceptual reproduction of spatial sound using loudspeaker-signal-domain parametrization," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 23, no. 10, pp. 1660–1669, 2015.

[27] S. Delikaris-Manias, J. Vilkamo, and V. Pulkki, "Parametric binaural rendering utilizing compact microphone arrays," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 629–633, IEEE, 2015.

[28] J. Vilkamo, T. Bäckström, and A. Kuntz, "Optimized covariance domain framework for time–frequency processing of spatial audio," *Journal of the Audio Engineering Society*, vol. 61, no. 6, pp. 403–411, 2013.

[29] D. P. Jarrett, E. A. Habets, and P. A. Naylor, "3d source localization in the spherical harmonic domain using a pseudointensity vector," in *2010 18th European Signal Processing Conference*, pp. 442–446, IEEE, 2010.

[30] J. Ivanic and K. Ruedenberg, "Rotation matrices for real spherical harmonics. direct determination by recursion," *The Journal of Physical Chemistry A*, vol. 102, no. 45, pp. 9099–9100, 1998.

doi:10.25836/sasp.2019.26

INTER-FREQUENCY BAND CORRELATIONS IN AUDITORY FILTERED MEDIAN PLANE HRTFS

Yukio Iwaya[1] **Brian FG Katz**[2] **Tetsu Magariyachi**[3] **Yôiti Suzuki**[4]
[1] Tohoku Gakuin University, Japan & Sorbonne Université, France
[2] Sorbonne Université, CNRS, Institut Jean Le Rond d'Alembert, France
[3] Tohoku University, Japan (Currently SONY Corp., Japan)
[4] Tohoku University, Japan
iwaya@ieee.org

ABSTRACT

In this paper, head-related transfer functions (HRTFs) in the median plane were analyzed to investigate details of spectral cues of sound localization in the median plane. Head-related impulse responses (HRIRs), which are representation of HRTFs in time domain, were filtered with auditory filters based on equivalent rectangular bands (ERB). Level changes according to median plane angle were calculated for each band and inter-band correlations were analyzed. Results showed some ERB bands, which include characteristic notches, are negatively correlated to other bands. Furthermore, whole frequency bands were grouped into three aggregated bands and their level changes were analyzed to clarify causes of negative correlations.

1. INTRODUCTION

Spectral cues in head-related transfer functions (HRTF), such as peaks and notches occurring above 4 kHz, are important for sound localization in the median plane [1–5]. However, it may be complicated for the auditory system to detect absolute frequency and level peaks and notches, mapping them to three-dimensional positions. In contrast, it may be more reasonable that comparisons are made of the relative level differences between frequency bands due to various peaks and notches. With this approach, it is not necessary to detect peaks and notches directly, only comparisons in levels across frequency bands are needed. In this paper, we analyze level changes of median plane HRTFs in narrow frequency bands using auditory filters and inter-band correlations. These changes are investigated to clarify effects of peaks and notches on comprehensive level changes in the corresponding HRTFs.

© Yukio Iwaya, Brain FG Katz, Tetsu Magariyachi, and Yôiti Suzuki. Licensed under a Creative Commons Attribution 4.0 International License (CC BY 4.0). **Attribution:** Yukio Iwaya, Brain FG Katz, Tetsu Magariyachi, and Yôiti Suzuki. "Inter-frequency band correlations in auditory filtered median plane HRTFs", 1st EAA Spatial Audio Signal Processing Symposium, Paris, France, 2019.

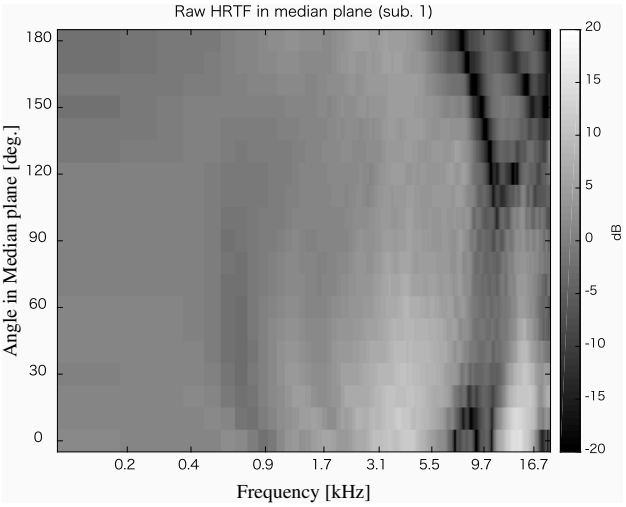


Figure 1. An example of HRTFs in median plane (HRTF 1).

2. OVERVIEW OF HRTFS FILTERED BY AUDITORY FILTERS

We analyzed 105 HRTF sets from the RIEC (Research Institution of Electrical Communication, Tohoku University) database, available in the SOFA [6] format standard. HRTFs were measured using a spherical loudspeaker array at RIEC for individual listeners. Head-related impulse responses (HRIRs) were acquired in the upper median plane from front (0°) to rear (180°) in 10°-steps. Fig. 1 shows an example median plane HRTF. We can observe deep notches from 4 kHz that shift with changed in median plane angle. These notches are considered to be due to acoustic phenomena of the pinnae. Furthermore, shallow notches can be seen at low frequencies. These notches are associated to the head being a rigid obstacle.

Each HRIR was then filtered by a band limited auditory filter. A Gammatone filter bank [7] was employed in this analysis, with 40 equivalent rectangular bandwidth (ERB) [8] over the full audible frequency range (up to 20 kHz). Output power level of the filtered HRIRs for the 19 median plane angles was calculated, resulting in 760 values (19 angles × 40 bands) for each listener. Gammatone filter on the time domain can be expressed as follows:

doi:10.25836/sasp.2019.33

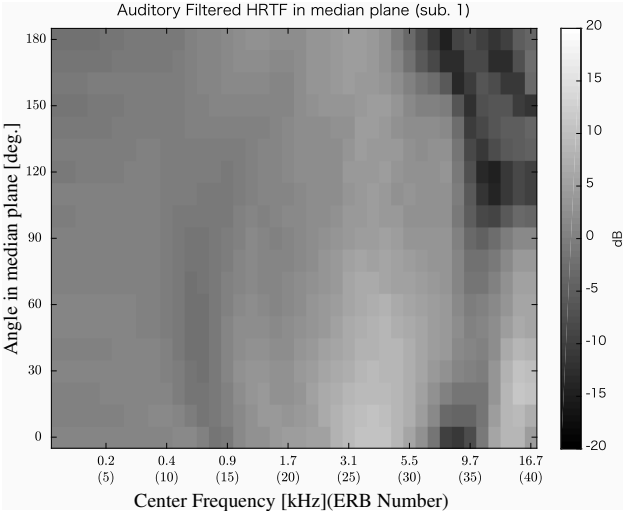


Figure 2. An example filtered median plane HRTF (HRTF 1). Numbers in parentheses indicate ERB frequency band (ID#), numbered from 1 to 40 from the lowest frequency.

$$g(f_{cj}, t) = at^{n-1} \exp(-2\pi b \mathrm{ERB}(f_{cj})t) \cos(2\pi f_{cj} t + \phi),$$

where f_{cj} is the center frequency, and parameters in $g(f_{cj}, t)$ were set as follows:

$$\begin{aligned} a &= 1 \\ b &= 0.9826 \\ \phi &= 0 \\ n &= 10 \log_{10} 2. \end{aligned}$$

Fig. 2 shows an example of a filtered median plane HRTF. Although the frequency resolution is reduced due to the use of band-passed filters, frequency notches can still be observed.

3. CORRELATION ANALYSIS

From the output levels of each ERB band, the level change of individual ERB bands was obtained as a function of elevation angle: $P(\theta_k, f_{ci}), (\theta_k = 0, 10, ..., 180°)$. The correlation across frequency bands for the level change as a function of angle was then calculated. The correlation coefficient value $C(f_{ci}, f_{cj})$ is expressed as:

$$C(f_{ci}, f_{cj}) = \frac{\sum_{k=1}^{N} \left(P(\theta_k, f_{ci}) - \overline{P(\theta, f_{ci})}\right) \left(P(\theta_k, f_{cj}) - \overline{P(\theta, f_{cj})}\right)}{\sqrt{\sum_{k=1}^{N} \left(P(\theta_k, f_{ci}) - \overline{P(\theta, f_{ci})}\right)^2 \sum_{k=1}^{N} \left(P(\theta_k, f_{cj}) - \overline{P(\theta, f_{cj})}\right)^2}}$$

This produced 39 cross-correlation values and one auto-correlation for each band with a correlation matrix of 40 bands × 40 bands for each listener. Fig. 3 shows an example correlation matrix for an HRTF. The diagonal expresses the auto-correlation coefficient of the band, equal to unity. Fig. 3 is symmetric with respect to the diagonal. It can be observed that there are ERB-pairs with negative correlations. Level changes of these ERB-pairs are opposite each other. These tendencies were observed in all 150 HRTF sets. There are two frequency band regions which have negative correlations with respect to the remaining frequency bands. These are the frequency ranges of $\approx 0.4 - 1$ kHz and $\approx 6 - 10$ kHz.

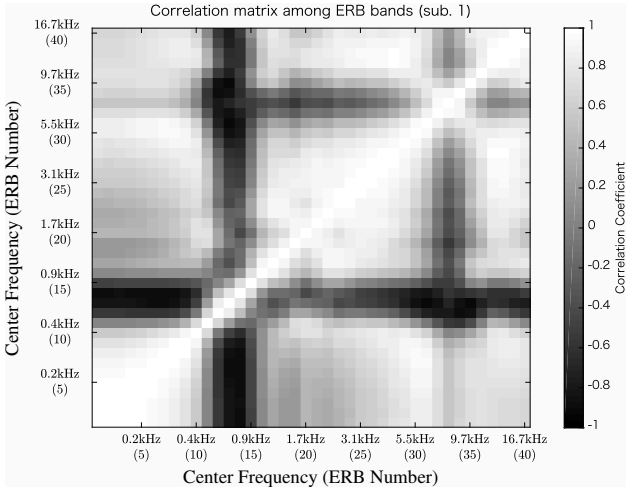


Figure 3. An example ERB band correlation matrix (HRTF 1).

4. ANALYSIS OF ERB BAND LEVEL CHANGES

The correlation coefficient cannot represent the absolute value of the level change. Therefore small change could be emphasized in previous analysis in Sec 3. We confirmed concrete level changes of all bands. Fig. 4 shows the actual level change of each band as a function angle. Observed level changes for low bands (ERB bands #1–#15) in the first sub-figure are small (2–3 dB), less than observed in the other bands. We can find the levels (ERB bands #11–#15) are slightly increasing in the first sub-figure of Fig.4. In contrast, the level changes in second and fourth sub-figures tends to decrease. Therefore, negative correlation would be obtained and emphasized nevertheless level change is very small in low frequency bands.

In the second sub-figure (ERB bands #16–#30), level changes tend to decrease. In the fourth sub-figure (ERB bands #37–#40) shows similar tendencies. Level change in the third sub-figure, however, relative level rises then falls. This is due to the frequency shift of the characteristic notches.

From these results, since the change of the lower frequency bands having a strong negative correlation is small, the influence on perception is considered to be small. In contrast, the level change in the bands in the third sub-figure, which have a negative correlation, is characteristic because their level changes are large and dynamic.

In the next section, based on these analysis results, the

doi:10.25836/sasp.2019.33

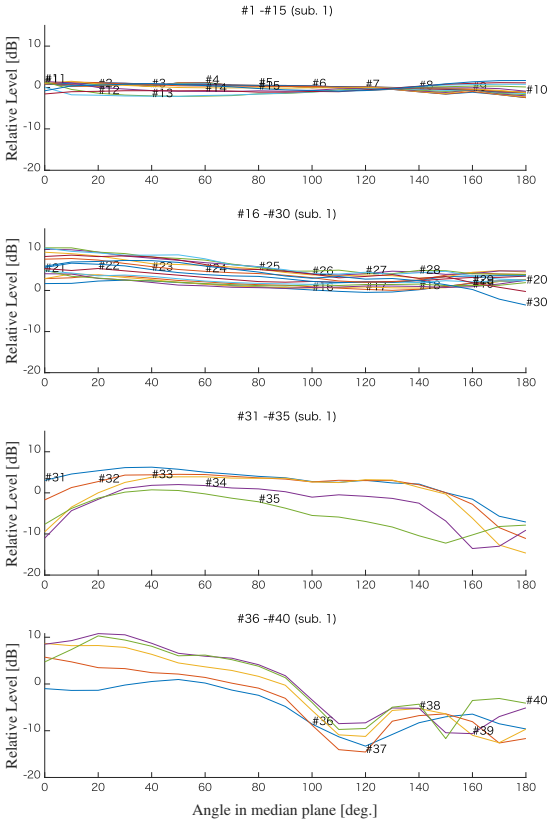


Figure 4. Level change of each ERB band as a function of median plane angle in Fig. 2. ID# indicates the ERB number (HRTF 1).

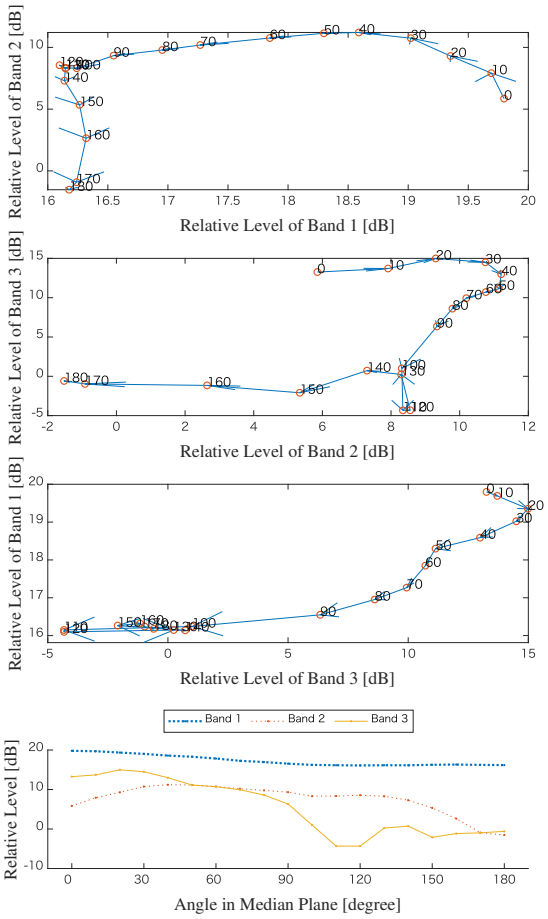


Figure 5. Scatter plot of progression of relative level between grouped bands. ID#'s along plot line indicate median plane angle. Data points are connected by arrows according to median plane angle progression. The bottom sub-figure shows relative level of grouped bands according to median plane angle (HRTF 1).

ERB bands are grouped into three regions and their relative changes are analyzed to clarify the negative correlation bands, including the characteristic notches.

5. ORIGIN OF NEGATIVE CORRELATION IN FREQUENCY REGION INCLUDING CHARACTERISTIC NOTCHES

Although there are two frequency regions with negative correlation values, the absolute level changes for the lower region (ERB bands #11–#15) is small and therefore ignored in subsequent analysis. In terms of the auditory system, it is assumed that it is reasonable to group frequency bands, which have similar level changes in median plane. Therefore, we divide the audible frequency bands into three large frequency region bands:

1. Band 1 (ERB bands #1–#30), which includes first and second sub-figures of Fig 4.
2. Band 2 (ERB bands #31–#36), in which the characteristic notches are included
3. Band 3 (ERB bands #37–#40), in which level change tends to decrease in median plane as you can see in fourth sub-figure of Fig. 4.

Figure 5 shows the progression of relative level as a function of angle between the three region bands. The forth figure shows relative level of three grouped bands along to angle in median plane.

From Fig. 5, considering Band 1 & Band 2, it can be seen that there are opposite changes (levels decreasing in one band while increasing for the other) from $0°$ to $40°$ in front, and from $140°$ to $180°$ at the rear. The negative correlation is caused from these tendency. In contrast, from $50°$ to $130°$, the level progression is positive.

From the second sub-figure (Band 2 & Band 3), the relation from $150°$ to $180°$ is slightly negative. In front, this negative relation can be partially observed from $20°$ to $50°$. From the third sub-figure (Band 1 & Band 3), the relationship is almost always positive, because relative levels in both Band 1 and Band 3 decrease with median plane angle. The negative correlation observed for Band 2 is due to the frequency shifting of the characteristic notches.

doi:10.25836/sasp.2019.33

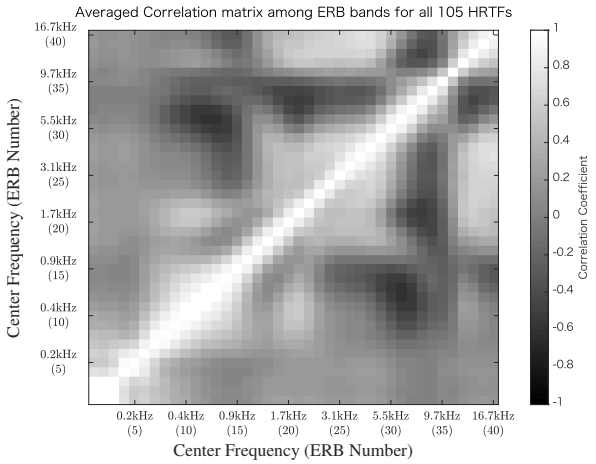


Figure 6. Correlation matrix averaged across all 105 HRTFs.

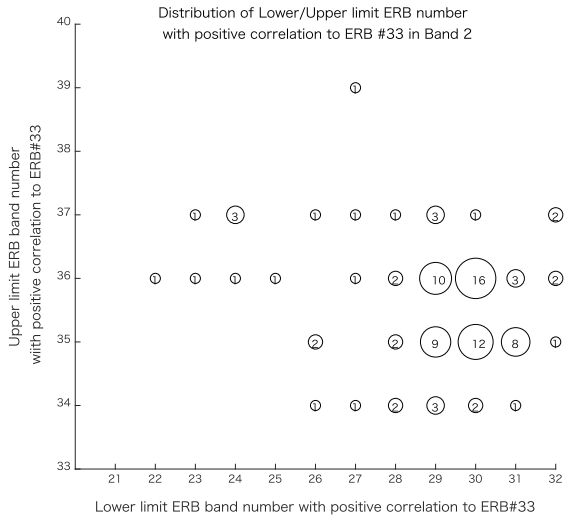


Figure 7. Distribution of upper/lower ERB band number with positive correlation to ERB #33 for all 150 HRTF sets

Fig. 6 shows the correlation matrix averaged across all 105 HRTFs. From Fig. 6, the negative correlation of level change in the grouped Band 2, in which the characteristic notch exists, was still observed. Therefore, this negative correlation in Band 2 is common in HRTF spectrum.

Furthermore, we investigated details of band width of Band 2. Fig. 7 shows distribution of upper and lower limit ERB band number of Band 2 with positive correlation to ERB #33, which is included in the grouped Band 2, for all HRTFs. Number in each circle indicates number of HRTF sets in the condition. We can see that the spread of positive correlations is different among HRTFs. In other words, the band width of Band 2 differs among individuals. These are caused from the fact that the frequency shift of characteristic notches differs among individuals.

6. CONCLUSION

In this paper, upper median plan HRTFs were analyzed and inter-band correlations were investigated to examine relationships between spectral cues associated with sound localization in the median plane. Results showed that the audible frequency could be divided into three frequency bands. The middle band, comprising the characteristic pinnae notches, showed a progression of of level as a function of elevation angle which had a negative correlation relative to the upper and lower bands.

It is hypothesized that this relative information could be exploited by the auditory system to provide elevation information. If shown in perceptual studies, this could indicate that we do not need to detect frequency notches directly. Namely, we may compare relative level across three large frequency bands. Therefore, listening test are the next step in our work to investigate this hypothesis.

7. ACKNOWLEDGEMENTS

This study was partially supported by JSPS KAKENHI (19K12286, 19001004).

8. REFERENCES

[1] K. Iida, M. Itoh, A. Itagaki, and M. Morimoto, "Median plane localization using a parametricmodel of the head-related transfer functionbased on spectral cues," *Applied Acoustics*, vol. 68, pp. 835–850, 2007.

[2] M. Morimoto and H. Aokata, "Localization cues of sound sources in the upper hemisphere," *J Acoust Soc Jpn (E)*, vol. 5, no. 3, pp. 165–173, 1984.

[3] J. Hebrank and D. Wright, "Spectral cues used in the localisation of sound sources on the median plane," *Journal of the Acoustical Society of America*, vol. 56, pp. 1829–1834, 1974.

[4] R. Greff and B. Katz, "Perceptual evaluation of HRTF notches versus peaks for vertical localisation," in *Intl. Cong. on Acoustics 19*, (Madrid, Spain), pp. 1–6, 2007.

[5] J. Blauert, *Spatial hearing: The Psychophysics of Human Sound Localization*. Cambridge, Massachusetts, US: The MIT Press, 1996.

[6] P. Majdak, Y. Iwaya, *et al.*, "Spatially oriented format for acoustics: A data exchange format representing head-related transfer functions," in *roc. 2013 International AES Convention*, (Rome, Italy), pp. 12–13, 2013.

[7] R. Patterson, M. Allerhand, and C. Giguere, "Time-domain modeling of pe- ripheral auditory processing: A modular architecture and a software platform," *J. Acoust. Soc. Am.*, vol. 98, pp. 1890–1894, 1995.

[8] B. Glasberg and B. J. Moore, "Derivation of auditory filter shapes from notched-noise data," *Hearing Research*, vol. 47, pp. 103–138, 1990.

doi:10.25836/sasp.2019.33

AUTHOR INDEX

Abraham 143
Alary 43
Amengual Garí 121, 161
Aramaki 67
Arend 49
Baldev 1
Bara 55
Basu 1
Baumgartner 155
Bavu 127
Berge 25
Boren 143
Bouchara 55
Brimijoin 161
Dagar 167
Delikaris-Manias 173
Demontis 13
Derrien 67
Dufor 73
Engel 121
Fargeot 67
Farrugia 73
Garcia 127
Giller 61
Gillioz 49
Gros 73, 79
Grzyb 143
Guilbert 55
Guo 85
Habets 19
Halmrast 115
Hassager 161
Hegde 1, 167
Henry 121
Herzog 19
Höldrich 61
Iwaya 179
Jiang 91
Katz 179
Kearney 97, 149
Klein 37
Kronland-Martinet 67
Lane 143
Langrenne 127
Le Prado 79
Liu 91
Lopez-Lezcano 103, 109
Lu 85
Maestre 31
Magariyachi 179
Mahé 7
Majdak 155
Marchal 13
Marchand 7
Massé 43
McCormack 173
McKenzie 97, 149
Merceruio 143
Moreira 79
Murphy 97, 149
Naressi 143
Natkin 79
Nicol 73, 79
Noisternig 43
Okumura 133
Ollivier 13
Otani 133
Parseihian 67
Picinali 121
Poirier-Quinot 121
Pörschmann 49
Ragot 7
Robinson 121, 161
Routray 1
Rueff 73
Scavone 31
Shabtai 137
Smith 31
Suzuki 179
Tzirkel 137
Välimäki 43
Viaud-Delmon 79
Vorländer 37
Wang 85
Weiss 55
Wendt 61
Xie 91
Yu 85
Zhang 91

https://sorbonne-universite.fr

https://sasp2019.ircam.fr