Spatial Decomposition Method for Room Impulse Responses

SAKARI TERVO, JUKKA PÄTYNEN, ANTTI KUUSINEN
(sakari.tervo@aalto.fi) (jukka.pätynen@aalto.fi) (antti.kuusinen@aalto.fi)

AND TAPIO LOKKI, *AES Member*
(tapio.lokki@aalto.fi)

Department of Media Technology, Aalto University School of Science, FI-00076 Aalto

This paper presents a spatial encoding method for room impulse responses. The method is based on decomposing the spatial room impulse responses into a set of image-sources. The resulting image-sources can be used for room acoustics analysis and for multichannel convolution reverberation engines. The analysis method is applicable for any compact microphone array and the reproduction can be realized with any of the current spatial reproduction methods. Listening test experiments with simulated impulse responses show that the proposed method produces an auralization indistinguishable from the reference in the best case.

0 INTRODUCTION

Spatial sound encoding and reproduction techniques are important tools for room acoustics research [1,2]. For perceptual evaluation of room acoustics, a spatial room impulse response is first measured, then encoded for a multichannel loudspeaker reproduction system, and convolved with anechoic music. This process of reproducing spatial sound from a spatial room impulse response is illustrated in Fig. 1. The last part of this spatial sound reproduction process is typically called convolution reverb.

Previous research has presented several spatial encoding methods that can be applied for spatial impulse responses. The spatial encoding methods can be divided to three groups according to their aim. In the first group the aim is to reproduce the originally measured sound field over a certain area. These methods include, First-Order Ambisonics (1.OA), Higher-Order Ambisonics (HOA) [3], and Wave-Field Synthesis (WFS) [4,5]. In contrast, in the second group, the binaural reproduction methods, the intention is to reproduce the sound pressure correctly at listener's eardrums by recording the soundfield close or at the eardrum [6,7]. In the third group, the starting point is to analyze and reproduce some of the spatial cues correctly [8,9]. An example of an analysis method belonging to the third group is the Spatial Room Impulse Response Rendering (SIRR) [10]. The first two groups require specialized microphones or microphone arrays, whereas the methods of the last group aim to present signal processing schemes that are applicable, at least to some extent, for several microphone arrays. An advantage of the first two groups is that they can be applied to a continous signal, such as speech or music. It should be stated that this paper concentrates on the spatial encoding of the spatial room impulse responses, not continuous signals.

Most of the professionals working in the field will agree, that when applied in a careful manner, any of the aforementioned encoding techniques will provide a realistic or at least plausible auralization of the acoustics. However, the use of special microphone arrays imposes limitations to the measurement procedure. First, some of the microphones are known to have inaccurate directional response, especially in the higher frequencies, which naturally affects the accuracy of the analysis and the reproduction. Second, the measurement, especially for WFS, either requires a multitude of microphones or is time consuming, which can be costly. Finally, the microphone array setups for some of the reproduction approaches are only limited to that specific approach.

This paper presents a spatial encoding technique for spatial room impulse responses, named here Spatial Decomposition Method (SDM). In contrast to previously developed methods, SDM can be applied for an arbitrary compact microphone array with a small number of microphones and any spatial sound reproduction technique. The presented method relies upon the simple assumption that the sound propagation direction is the average of all the waves arriving to the microphone array at time *t*, and the sound pressure of a single impulse response in the geometric center of the array is associated with it. The method analyzes the spatial

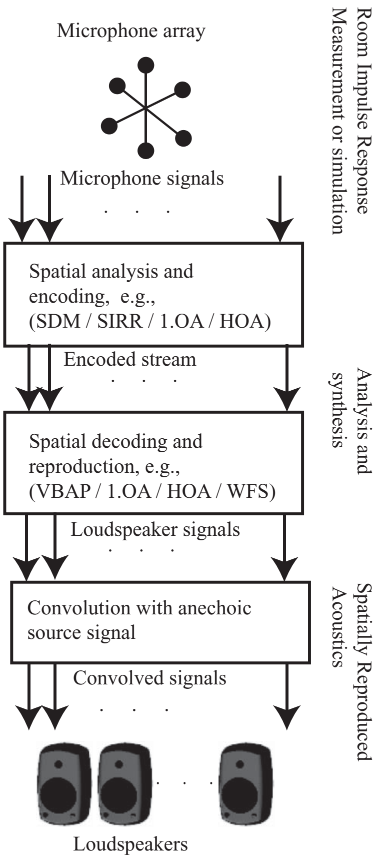

Microphone array
Microphone signals
Spatial analysis and encoding, e.g., (SDM / SIRR / 1.OA / HOA)
Encoded stream
Spatial decoding and reproduction, e.g., (VBAP / 1.OA / HOA / WFS)
Loudspeaker signals
Convolution with anechoic source signal
Convolved signals
Loudspeakers
Room Impulse Response Measurement or simulation
Analysis and synthesis
Spatially Reproduced Acoustics

Fig. 1. The general processing applied in auralization using room impulse response measurements or simulations. The acronyms for the encoding and decoding techniques are given in the text.

impulse response with this assumption and encodes it to a response that consists of samples that have a pressure value and a spatial location.

1 THEORY AND METHODS

This section presents theoretical background on the spatial room impulse response and the proposed spatial analysis.

1.1 Room Impulse Response

A room impulse responses captured with a microphones n at location $\boldsymbol{r}_n$ is the sum of individual acoustic events $h_{p,n}(t)$:

$$h_n(t) \overset{\triangle}{=} h(t|\boldsymbol{r}_n, \boldsymbol{x}) = \left[\sum_{p=0}^{P} h_{p,n}(t)\right] + w_n(t)$$
$$= \left[\sum_{l=0}^{l} \left(\int_{-\infty}^{\infty} H_{p,n}(\omega) e^{j\omega t} d\omega\right)\right] + w_n(t), \qquad (1)$$

where n denotes microphone index, t is time, ω is the angular frequency, $\boldsymbol{x}$ is the source position, $p = 0, \ldots, P$ is the index for each acoustic event, $w_n(t)$ is the measurement noise, and $H_{p,n}(\omega)$ is the frequency domain representation of $h_{p,n}(t)$. The acoustic events can be, for example, the direct sound, discrete reflections, diffractions, or diffuse reflections. At each time moment t, the sound pressure at receiving location $\boldsymbol{r}_n$ has a scalar value, i.e., it is a scalar function $h(\boldsymbol{r}_n, \boldsymbol{x}|t)$. The scalar value is the overall sum of different sound pressure waves arriving at the same time to the receiver location. In the context of this paper the spatial room impulse response is measured with $n = 1, \ldots, N$ microphones, i.e., a microphone array.

The whole impulse response is altered by several acoustic phenomena. A majority of the acoustic events is attenuated according to $1/r$-law and affected by air absorption. In addition, the frequency response of an event is altered by the absorption of the surfaces in the enclosure. Moreover, the directivities of the microphones and the sound source have an effect on the impulse response.

As time progresses, the number of acoustic events per time window increases. In room acoustics research and convolution reverberation engines, the impulse response is traditionally divided into three consecutive regions in time: the direct sound, the early reflections, and the late reverberation. Next subsections list features of these categories in theory and in practice.

1.1.1 Direct Sound, Specular and Diffuse Reflections, and Diffraction

In theory, the direct sound is a single impulse, i.e., a Dirac delta function. Moreover, with the assumption of ideal reflecting surfaces, the reflections are also impulses. In practice, ideal specular reflections are rare, since they require an infinite rigid and flat plane. Thus, the early reflections are often spread over time instead of being single events in the impulse response and have a certain frequency response due to the absorption at the boundaries. Also, due to the non-ideal response of the loudspeakers and microphones, the direct sound is not an impulse. Moreover, the loudspeaker impulse response is typically different in all directions, therefore, reflections from different directions have different responses in time and frequency.

The concept of image-source describes an ideally specular reflection from a surface [11]. Although such reflections are rare or not even possible in real situations, the model can be used to describe several acoustic events. First, the diffraction from an edge can be modeled with properly weighted image-sources [12]. Second, non-ideal reflections, i.e., diffuse reflections, are caused by diffraction in a very small scale [13]. Third, close-to-ideal specular reflections and the direct sound can be modeled with a limited number of properly weighted image-sources and a source, respectively. Thus, it can be concluded that the acoustics of an enclosure can be modeled to some extent with a limited number of image-sources. However, in practice, the acoustic modeling of complex room geometries with image-sources is a very demanding task.

1.1.2 Late Reverberation

The reflection density increases in the room impulse response as the time progresses. When enough reflections arrive during the same time, the sound field becomes

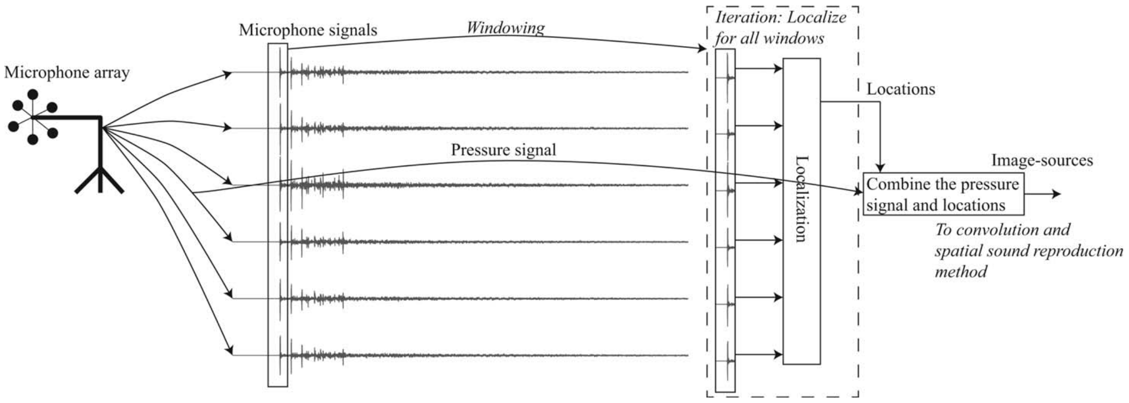

Microphone array
Microphone signals
Windowing
Iteration: Localize for all windows
Locations
Pressure signal
Localization
Image-sources
Combine the pressure signal and locations
To convolution and spatial sound reproduction method

Fig. 2. The processing in the proposed spatial encoding method consists of localization and combining the omni-directional pressure signal with the estimated locations.

diffuse. A diffuse sound field is spatially homogeneous and isotropic. In practice, this means that the distributions of the phase and direction are uniform and amplitude is equally distributed for each position. It follows from these conditions that the net energy flow over a volume is zero. The time when this occurs in an impulse response is typically referred to as the mixing time [14], and after that the impulse response is considered to be late reverberation.

1.2 Analysis

SDM assumes that the impulse response can be presented as a set of limited number of image-sources. SDM analyzes the spatial room impulse response at every discrete time step $\Delta t = 1/f_s$, where f_s is the sampling frequency. The sound arriving during these time windows has an average direction that is estimated with robust localization methods. As a result, a set of discrete pressure values and their corresponding locations, i.e., image-sources, present the spatial room impulse response. Decomposition of the image-sources describes this process and thus the name SDM (Spatial Decomposition Method). The overall processing in the method is illustrated in Fig. 2.

The analysis assumes the following general requirements for the used microphone array:

- For 3-D spatial sound encoding, the minimum requirement of the number of microphones is four, which are not on the same plane, so that they can set up a 3-D space.
- The directivity of one of the microphones is omnidirectional or it is possible to create one virtual omnidirectional pressure microphone signal from the others.
- The dimensions of the array are not large, i.e., the microphone array is compact. The dimensions should be less or equal to the dimensions of a human head.
- Open microphone arrays are preferred, but closed ones can also be used as long as the above requirements are met.

In detail, for a set of room impulse responses $\boldsymbol{H}(t) = \{h_n(t)\}_{n=1}^{N}$, i.e., a spatial room impulse response, the analysis proceeds as follows.

1.2.1 Step 1: Localization

First, SDM solves the location of the source and image-sources from the spatial room impulse response. For each discrete time step, a localization function $P(\cdot \mid \cdot)$ estimates the average location of the arriving sound in a small time window with respect to the geometric center of the array. The localization function maps the received data into a cost function that is given for a location $\boldsymbol{x}$ and possible parameters or a priori models $\boldsymbol{\chi}$:

$$\hat{\boldsymbol{x}}_k = \arg\max_{\boldsymbol{x}} \{P(\boldsymbol{H}(\boldsymbol{k})|\boldsymbol{x}, \boldsymbol{\chi})\}, \tag{2}$$

where $\boldsymbol{H}(\boldsymbol{k})$ is the spatial impulse response in a short time window that is defined by vector $\boldsymbol{k} = [-L/2 + k \ldots L/2 + k]\Delta t$ with discrete time indices at time Δtk, $k = 1 \ldots K$, where K is the length of the impulse response, and window size L.

The a priori models and the localization function depends on the applied microphone array, measurement conditions, and assumption on the sound field propagation model. As an example for an arbitrary array with arbitrary directivities, one can apply the maximum likelihood estimation given in [15] with the reverberation parameter set to $\gamma = 0$. For acoustic vector-sensors, e.g., a gradient microphone array, one can apply the solutions given in [16] or [17]. Moreover, [18] gives an overview of different localization functions that are based on time difference of arrival and time of arrival estimation. The accuracy of the localization depends on the applied microphone array and localization method, as well as on the conditions during the measurements.

This paper uses the least squares solution for time difference of arrival estimates (TDOA) for localizing the image-sources. Plane-wave propagation model is assumed for the localization since an efficient estimator for the problem exists, unlike for spherical wave propagation model [19] and since the source and the image-sources can be assumed to be in the far field. Although the solution is a set of plane-waves instead of a set of image-sources, the method can treat them in a similar manner due to the far-field assumption.

The TDOAs are obtained from generalized correlation method with direct weighting [20] for each time step in the small analysis window k. In addition, each TDOA estimate is interpolated with the exponential fit [21]. The TDOA estimates are denoted with

$$\hat{\boldsymbol{\tau}}_k = [\hat{\tau}_{1,2}^{(k)}, \hat{\tau}_{1,3}^{(k)}, \ldots, \hat{\tau}_{N-1,N}^{(k)}]^{\mathrm{T}},$$

where N is the number of microphones, and the corresponding microphone position difference vectors are denoted with

$$\boldsymbol{V} = [\boldsymbol{r}_1 - \boldsymbol{r}_2, \boldsymbol{r}_1 - \boldsymbol{r}_3, \ldots, \boldsymbol{r}_{N-1} - \boldsymbol{r}_N]^{\mathrm{T}}.$$

The least squares solution for slowness vector is then given as [22, p. 75]:

$$\hat{\boldsymbol{m}}_k = \boldsymbol{V}^{+}\hat{\boldsymbol{\tau}}_k, \tag{3}$$

where $(\cdot)^{+}$ is Moore-Penrose pseudo-inverse, and the direction of the arriving sound wave is given as $\hat{\boldsymbol{n}}_k = -\hat{\boldsymbol{m}}_k/\|\hat{\boldsymbol{m}}_k\|$. The distance to the image-source k is given directly by the time index and the speed of sound $d_k = ck\Delta t$.

1.2.2 Step 2: Dividing the Omni-Directional Pressure Signal

The second step of the analysis selects one of the available omni-directional microphone signals as the pressure signal h_p. Ideally, the microphone for the pressure signal is located in the geometric center of the array. In this case, the analysis assigns each sample of the pressure impulse response $h_p(\Delta tk)$ with a 3-D location $\hat{\boldsymbol{x}}_k$, which is the output from Step 1. Then, the method has encoded the spatial impulse response with four values per sample, the pressure value and the 3-D location of the sample.

In case the pressure microphone is not in the geometric center of the array, one has to predict the value of the pressure signal according to the image-source locations. This is done by first calculating the distance from the image-source location to the location of the pressure microphone $\boldsymbol{r}_p$

$$d_k = \|\boldsymbol{r}_p - \boldsymbol{x}_k\|, \tag{4}$$

and then assigning each image-source with the pressure value $h_p(f_s d_k/c)$. When using plane wave propagation model, the distance is calculated as

$$d_k = |\boldsymbol{n}_k(\boldsymbol{r}_p - \boldsymbol{x}_{k,0})|, \tag{5}$$

where $\boldsymbol{n}_k$ and $\boldsymbol{x}_{k,0}$ are the plane normal and a point on the plane, respectively.

Instead of predicting the pressure in the center of the array, one can predict the image-source locations in the location of the pressure signal. This is an easier choice because it does not require resampling of the signal. This paper applies neither of these approaches, since the pressure microphone is always located in the middle.

1.3 Limitations on the Performance and the Effect of the Window Size

Several aspects affect the accuracy of the analysis in SDM. When the noise level decreases and the number of microphones increases, the performance of the localization improves, as predicted by the Cramér-Rao lower bound (CRLB) (see, e.g., [18]). Other important factors are the time interval between the samples (Δt) and the size or the dimensions of the microphone array. The smaller these values are the more spatial and temporal separation between individual acoustic events can be made. This improves the localization for individual acoustic events. Other methods require larger aperture size to improve the approximations for low frequencies, however, in SDM this is not a requirement since in SDM the lower frequencies can be estimated by elongating the window size. However, this would also require that SDM processing is done for different frequency bands with different window sizes. This is further discussed in Section 4.1.

A limiting factor for the window size is the largest dimension of the microphone array. That is, the window size should be larger than the time that it takes for a sound wave to travel through the array, i.e., $L\Delta t > 2d_{\max}/c$, where $d_{\max}$ is the maximum distance between any two microphones in the array. Theoretically, a large window size improves and worsens the localization performance at the same time. Namely, as the window size increases, the localization performance of a single acoustic event improves, as stated by the CRLB. However, the probability that more than one acoustic event is present in the analysis window increases. This latter part is seen as a possible problem in the analysis and, therefore, it is recommended that the window size is selected such that it is just over $2d_{\max}c$. In addition, if an acoustic event is assumed to be short, time-wise, increasing the window size would actually decrease the theoretical performance since the energy of the noise in the time window increases relative to the energy of the signal, thus decreasing the signal-to-noise ratio.

The next part assesses the effect of the window size selection with a quantity called echo density. The echo density describes the average number of echoes in a room per a time instant and is valid for any arbitrarily shaped enclosure [23, p. 92]. It is defined as

$$\frac{N_r}{dt} = 4\pi\frac{c^3 t^2}{V}, \tag{6}$$

where N_r is the number of reflections, and V is volume. Echo density is a useful tool for inspecting the effects of the window size selection on the number of acoustic events, i.e., image-sources, per time window. The threshold when there is less than N_r reflection(s) present in the time window can be examined with

$$\tau_1 = \sqrt{\frac{N_r V}{dt 4\pi c^3}} \approx 0.0014\sqrt{V}. \tag{7}$$

The last approximation is yielded for less than one reflection $N_r \to 1$ and assuming that the speed of sound is constant $c = 345$ m/s. For example, a window size of $dt = L\Delta t = 1$ ms produces the value $\tau_1 = 119$ ms for a room with volume ($30 \times 20 \times 12 = 7200$ m^3), which indicates that there is only one acoustic event present in the analysis window until 119 ms after the direct sound. Thus, the parameter τ_1 describes the average time when there will be more than one reflection present in the analysis time window. The smaller the window size, the bigger the parameter τ_1 and the more accurate localization of individual acoustic events is achieved. To conclude, shorter time windows should be preferred over long ones, and the minimum length of the time window is defined by the maximum of the spacing between any microphone pair.

1.4 Rationale for SDM

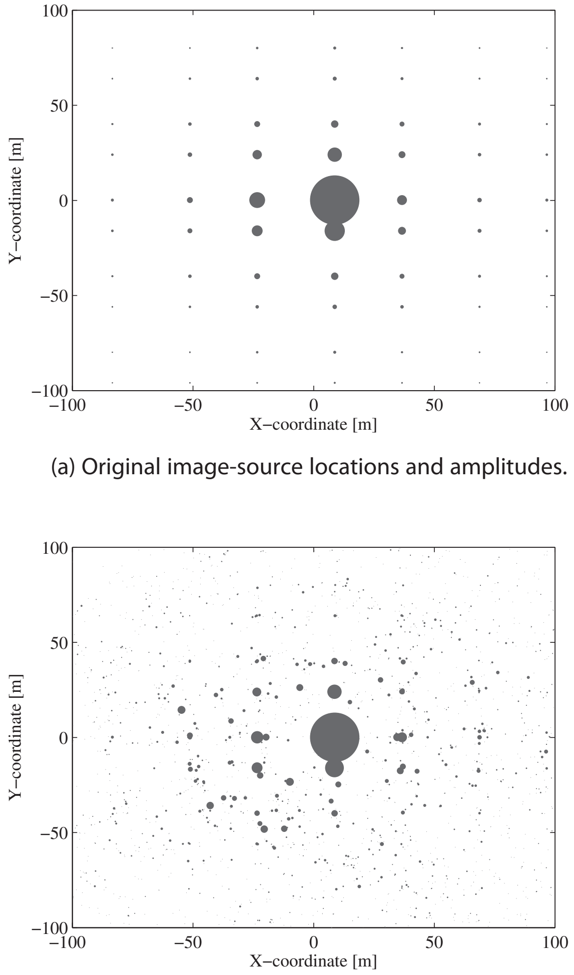
The accurate localization of first acoustic events with respect to time in the impulse responses, i.e., the direct sound and the first reflections, is possible as shown in [24] and [18], respectively. However, as the time progresses the number of acoustic events per time window increases, and eventually more than one reflection arrives during the time window. In this case, a cross-correlation-based localization algorithm localizes the sound to the location of the reflection that is the strongest one in that time window. The strongest direction is selected because it shows as the strongest peak in the cross-correlation functions. Analogous example of this behavior with one localization algorithm is shown with speech sources in [25]. However, it is also possible that the estimated location is an intermediate point that is between the reflections within that analysis window. This is, for example, the case if the localization algorithm is based on the average direction of the sound intensity. Thus, the estimated location depends highly on the localization algorithm. The behavior of the localization algorithms in the case of several acoustic events should be further investigated, but here this is left for future research. In any case, SDM assumes in the spatial reproduction that the estimated location corresponds to the correct perceptual location. The assumption has been used previously for example in SIRR [10,26].

SDM produces the diffuse sound field naturally. Namely, in SDM each time step has a random direction in a diffuse sound field. The total directional distribution over the total diffuse sound field, i.e., late reverberation is then uniform. Further evidence for this is provided in a recently published article which uses SDM for spatial analysis [27].

Since the first acoustic events are correctly localized from spatial room impulse response in the SDM framework, and these events are known to have a very prominent effect on the perception of spatial sound [1,28], the resulting auralization should be credible. Moreover, the late part of the spatial room impulse response will be naturally presented as diffuse by SDM because multiple arriving reflections will produce random directions.

1.5 An Example of the Analysis with SDM

This section demonstrates the principles in SDM with an illustration of analysis results of spatial room impulse response. The spatial room impulse response is recorded from a simulation of a shoebox room of size (20 × 30 × 12) m^3. Furthermore, the source was at [16.04, 8.06, 3.58] m and the receiver at [7.35, 7.92, 3.22] m. In addition, the applied window was 1.33 ms Hanning window and overlap between two consecutive windows is 99%. Speed of sound was set to $c = 345$ m/s, sampling frequency to $f_s = 48$ kHz, reflection coefficient to 0.85, and reflections up to 45th order were simulated.



(a) Original image-source locations and amplitudes.

(b) Analyzed image-source locations and amplitudes

Fig. 3. An example of the locations and amplitudes of (a) simulated image-sources and (b) decomposed image-sources with SDM from a spatial room impulse response. The area of each filled circle illustrates the energy of that image-source. The image-sources with the highest energy are correctly analyzed.

Fig. 3, where the radius of each circle corresponds to the amplitude of respective image-source, illustrates the results of the analysis. As can be seen in Fig. 3, the early part of the simulated spatial room impulse response (a) is very similar to the one analyzed by SDM (b).

2 LISTENING TEST EXPERIMENTS

This section describes the listening test setup, the listening room, the simulated room acoustic conditions, and the source signals. In addition, listening test procedures and results are presented.

This paper uses Vector Base Amplitude Panning (VBAP) [8] as the spatial reproduction technique for the listening tests. Other reproduction methods could also be used, but VBAP is here preferred since it can be implemented for a 3-D spatial sound with less number of loudspeakers than the other methods and since it provides good subjective quality in overall. The listening tests compares the proposed method to SIRR [10,26], which can be considered the

Table 1. Reverberation time (RT), sound pressure level (SPL), and noise level (NL) in the listening room. Sound pressure level is given with respect to the reference (Ref.) value at 200 Hz-4 kHz frequency band. In the calibration, the SPL was 87 dB, which gives a signal-to-noise ratio of more than 45 dB for each octave band.

Octave band	RT [s]	SPL [dB]	NL [dB]
Ref. [200 Hz - 4 kHz]	0.14	0.00	-
125 [Hz]	0.24	1.66	39.9
250 [Hz]	0.17	0.47	35.7
500 [Hz]	0.13	0.36	32.9
1 [kHz]	0.13	0.02	28.6
2 [kHz]	0.12	0.16	20.4
4 [kHz]	0.11	−1.03	18.9
8 [kHz]	0.10	−2.93	21.5

Table 2. The azimuth and elevation directions and distance of each individual loudspeaker (LPS) in the 14-channel loudspeaker reproduction setup in the listening room. 8, 4, and 2 loudspeakers are located approximately at the lateral plane, 45 degrees above lateral plane, and –45 degrees below lateral plane, respectively. The loudspeakers were localized using the method presented in [24].

LPS #	Azimuth [°]	Elevation [°]	Distance [m]
1	46.7	0.1	1.01
2	89.6	−1.5	1.02
3	134.2	−0.1	0.98
4	179.8	−0.2	0.95
5	−135.9	0.1	1.02
6	−91.6	0.2	0.94
7	−45.6	1.0	0.96
8	−0.3	1.6	0.98
9	45.9	43.5	1.30
10	135.1	40.5	1.33
11	−137.9	42.3	1.40
12	−45.4	46.4	1.33
13	24.0	−46.1	1.29
14	−19.6	−45.1	1.27

state-of-the-art spatial sound encoding method for spatial room impulse responses, at least for VBAP. SIRR also operates under the same assumption as SDM, that the binaural cues are produced correctly.

2.1 Listening Room Setup and Stimuli

Listening tests were conducted in an acoustically treated room with dimensions of ($x \times y \times z : 3.0 \times 5.1 \times 3.8$) m^3. Table 1 shows the reverberation time, sound pressure level, and noise level in the listening room. The listening room fulfills the recommendations given by ITU in [29], with the exceptions that the noise level fulfills the noise rating (NR) 30 requirement, whereas the recommendation is NR 15 and the listening distance is about 1.2 meters on average, whereas the recommendation is more than two meters.

The listening room includes a 3-D 14-channel loudspeaker setup, out of which 12 are of type Genelec 8030A, and two are of type Genelec 1029A loudspeakers. Table 2 gives the location of each loudspeaker with respect to the listening position at the origin (0,0,0) m. Each loudspeaker is calibrated so that they produce equal *A*-weighted sound pressure level with *slow* temporal averaging in the listening position for a band-pass filtered noise from 100 Hz to 5 kHz. Since the distance of the loudspeakers is not the same to the reference position for all loudspeakers, they are all delayed with digital signal processing so that each loudspeaker is at a virtual distance of 1.40 m.

Table 3. Source and receiver positions, source signals, dimensions of the rooms, and sample naming used in the listening test. Speed of sound was set to $c = 345$ m/s, sampling frequency to $f_s = 48$ kHz, reflection coefficient to 0.85, and reflections up to 45th order were simulated.

Sample	Source position			Receiver position		
(Signal)	x [m]	y [m]	z [m]	x [m]	y [m]	z [m]
			Large room (30 × 20 × 12) m^3			
A (Sp.)	16.04	8.06	3.58	7.35	7.92	3.22
B (Tr.)	17.44	12.81	2.88	2.64	13.48	3.72
C (Ca.)	20.37	11.99	2.52	3.10	12.10	2.86
			Small room (5 × 3 × 2.8) m^3			
D (Sp.)	3.44	0.80	1.53	1.02	0.64	1.40
E (Tr.)	3.87	1.45	1.65	0.76	1.39	1.33
F (Ca.)	3.78	0.85	1.81	1.24	0.97	2.07

Sp.: Speech, Tr. Trombone, and Ca.: Castanet

The simulated impulse responses for the listening test were produced with the image-source method [11] in two modelled rectangular rooms. In the image-source method, throughout this paper, the reflection coefficient is set to 0.85, the speed of sound to 345 m/s, the sampling frequency to 48 kHz, and reflections up to 45th order are simulated. In addition, Table 3 shows the room dimensions, source, and receiver positions used in the image-source method. Two shoebox rooms, a large and a small one, are simulated for the listening tests. The large and the small room have wide band reverberation times of 2.0 s and 0.4 s, respectively. In all the cases, the room impulse responses are truncated from –40 dB onwards according to the backward integrated Schroeder curve.

2.1.1 Reference and Anchor

The reference was generated with the image-source method. The location and amplitude of each image-source was transferred into a virtual source, which was panned with VBAP for the current loudspeaker setup [8]. Finally, to simulate a real room impulse response measurement situation, the anechoic impulse response of Genelec 1029A was convolved with the impulse responses, and the impulse response was filtered with air absorption filters, implemented according to [30].

The anchor for the listening test was selected to be the same mono impulse response as in the reference but instead of VBAP processing according to the directional information obtained from the image-source method, it was used directly in the front loudspeaker (# 8 in Table 2).

2.1.2 Spatial Encoding Methods

Similarly to the reference, the image-source method was used to generate spatial impulse responses for a virtual

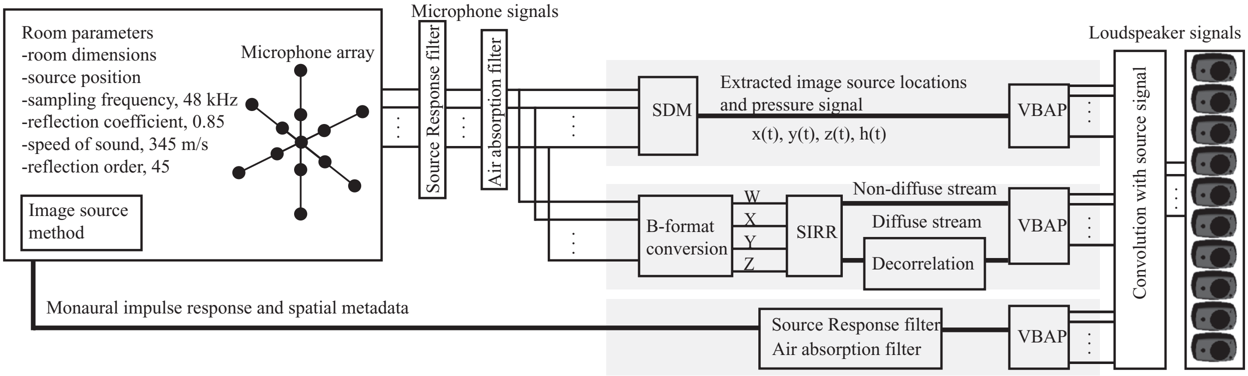


Fig. 4. Processing of the samples in the listening test experiments. The shaded areas highlight the different spatial encoding methods (from top to down, SDM, SIRR and reference).

Table 4. Origin centered coordinates for the microphone arrays. Spacing d_{spc} is equal for each microphone pair on a single axis.

Microphone #	X [m]	Y [m]	Z [m]
1	$d_{spc}/2$	0	0
2	$-d_{spc}/2$	0	0
3	0	$d_{spc}/2$	0
4	0	$-d_{spc}/2$	0
5	0	0	$d_{spc}/2$
6	0	0	$-d_{spc}/2$
7	0	0	0

microphone array. The microphone array consists of seven microphones, of which six are on a sphere and one in the geometric center of the array, as shown in Table 4. The central microphone is used as the microphone for the pressure signal in the spatial encoding methods.

The proposed spatial encoding method, SDM, was compared to two versions of SIRR. The first version of SIRR, as well as SDM, was implemented with seven microphones, and the second version of SIRR was implemented with 13 microphones. Their naming is the following:

- SDM with a single microphone array with spacing $d_{spc} =$ 100 mm and one microphone in the geometric center is named SDML7,
- SIRR with a single microphone array with spacing $d_{spc} = 100$ mm and one microphone in the geometric center is named SIRRL7, and
- SIRR with two microphone arrays with spacings $d_{spc} =$ 100 mm and $d_{spc} = 25$ mm and one microphone in the geometric center is named SIRR13.

The microphone arrays were selected as such, since SIRR-processing can be implemented for them [10,26]. Namely, SIRR requires the three components of particle velocity, which can be calculated with the gradient microphone-technique and a pressure signal, which is the microphone in the geometric center.

SIRR13 analyzes separately the room impulse responses for large and small spacing and in the post-processing phase combines them. Combination adds the analysis result for low frequencies below 1 kHz from the large spacer, and for high frequencies above 1 kHz with the smaller spacer. Before the addition, the analyzed signals for small and large spacer are low-pass and high-pass filtered with a 10th order Butterworth IIR filter, respectively. The motivation for such processing is that the present authors have used such array in measurements of concert halls [31].

To compare the methods in the same conditions, all the analyses use a Hanning window of 1.33 ms (64 samples at 48 kHz). SIRR has an overlap of 50% between two consecutive windows, and SDM has an overlap of 99% (63 samples), as explained in Section 2.2 (Step 1). The window size was selected as 1.33 ms since it is the one used in the original SIRR paper [26]. It should be emphasized that for SDM the optimal window size is much smaller than the selected one. Especially for the smaller simulated room, it is expected that the lengthy time window causes problem in SDM, since parameter τ_1 is 1.4 ms. However, since the goal is to compare these two techniques in the same conditions, the same window size is used for both. Moreover, a virtual microphone-based synthesis, originally developed for DirAC in [32], was noticed to provide a more natural sound for SIRR and was included in the processing.

The output of the SDM, i.e., the extracted image-sources, are directly panned with VBAP for the current loudspeaker setup. The output of the SIRR-analysis is processed as described in [10] and [26] for VBAP reproduction. In addition, the diffuse part of the SIRR is implemented with the Hybrid Method described in [26]. The processing of the listening test samples for SDM, SIRR, and the reference case are illustrated in Fig. 4.

2.1.3 Source Signals and Test Samples

Approximately ten seconds of male speech, trombone, and castanets were selected as the source signals. Each sample was convolved separately with the corresponding 14-channel VBAP output for a reference, SIRR, or SDM. The test samples are named from A to F as indicated in Table 3.

2.2 Listening Test Procedure

The task in the listening test is to compare the “similarity” of the spatially encoded samples with the reference sample,

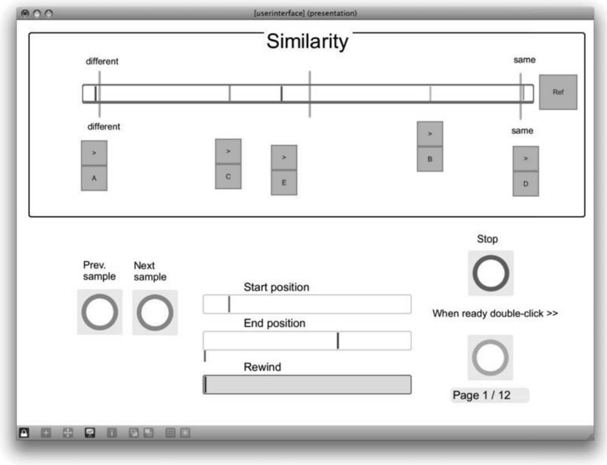


Fig. 5. Screen capture of the user interface (UI) used in the listening tests. The subjects can freely move, listen to, and rate the samples. Note that in the UI the alphabet stands for "method."

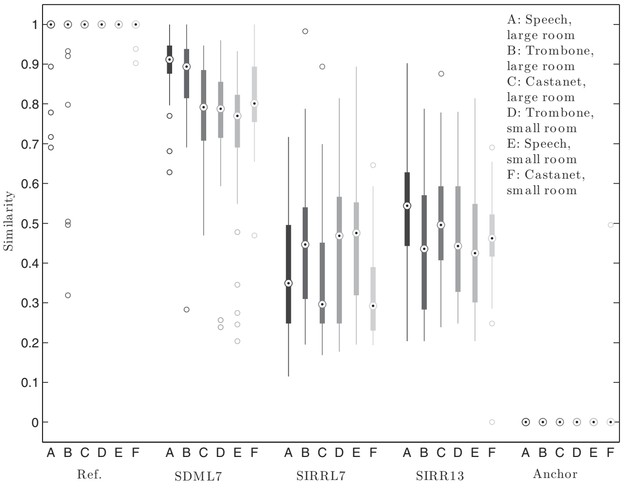


Fig. 6. Listening test results, the thicker boxes with solid color illustrate the 25 and 75 percentiles, the thinner lines illustrate the most extreme data points, the circles outside the boxes illustrate outliers, and the dots the median.

instead of "impairment" recommended by ITU [29]. This deviation from the ITU-recommendation was made since the test subjects are not encouraged to think that the samples are somehow impaired. In addition, the ITU-recommended impairment scale (imperceptible; perceptible, but not annoying; slightly annoying; annoying; very annoying) is not used in the listening test, since it is known from previous research [26] that SIRR-processed samples sound quite natural and are quite similar to the reference. Thus, the idea of this listening test is to find out which encoding method produces the sound that is most similar with the reference.

The listening test was implemented as a parallel comparison with continuous scale, and the task was to compare the similarity of five samples to a reference sample. Test subjects completed the test twice. The order of the test cases A – F (Table 3) was randomized between subjects and repetitions. Also, in each test case, the five samples (Ref., Anchor, SIRR13, SIRR7, SDML7) were presented in a random order with letters A – E. A screen shot of the user-interface and one comparative evaluation of one test case is shown in Fig. 5. During the listening test, the subjects could freely loop a time window what they were listening to and listen to an unlimited number of times. That is, there was no time limit for completing the test.

In the beginning of the test, the subjects had an adequate time to familiarize themselves with the samples. The test subjects were instructed to carefully consider the timbral and the spatial aspects in the samples. They were also told that one of the five samples is the hidden reference sample and one other sample is a mono anchor sample, which is played back from the front loudspeaker. After the familiarization, the subjects rated the samples according to similarity to the reference in the actual listening test. When the test ended, the subjects were interviewed and asked for the attributes that they used for discriminating and rating the samples.

Seventeen test subjects with normal hearing participated in the test. None of the subjects were the present authors of this paper. Most (9/17) of the test subjects can be considered expert listeners in spatial audio due to their background in spatial audio research. Others (3/17) had experience on critical listening, but this was not necessarily on spatial audio. These subjects were considered as experienced listeners. The rest (5/17) were naïve test subjects and had limited or no experience in critical listening. The test took on average approximately 50 minutes, including approximately a 10-minute familiarization step and a 5-minute interview.

3 RESULTS

All the results from the listening tests are shown in Fig. 6. The results from the listening tests are normalized for each test case and subject between 0 and 1. As shown in Fig. 6, the references and anchors are found correctly in most of the cases. SDML7 is mistaken as the reference 12 times, and the SIRR13 as the anchor once. This result already suggests that SDML7 is well suited for spatial encoding.

Multi-way analysis of variance (ANOVA) is applied to examine the main effects, and two- and three-factor interactions. The examined main effects are the spatial encoding method (Method), repetition of the test case (Repetition), the size of the room (Room), and the source sound sample (Sound). To perform ANOVA, the cases should be independent, the variances equal (homoscedasticity), and the residuals normally distributed. Here, the cases are assumed to be independent but other assumptions for ANOVA, the homoscedescacity and the normality of the residuals, are next tested with statistical tests.

Levene's test [33] shows that the variances between different test cases are significantly different. In addition, Anderson-Darling test [34] indicates that the residuals are not normally distributed. Both of these results are most likely a consequence of the scale in the listening test. That is, as shown in Fig. 6, the results for the reference and anchor sound condition are negatively and positively skewed,

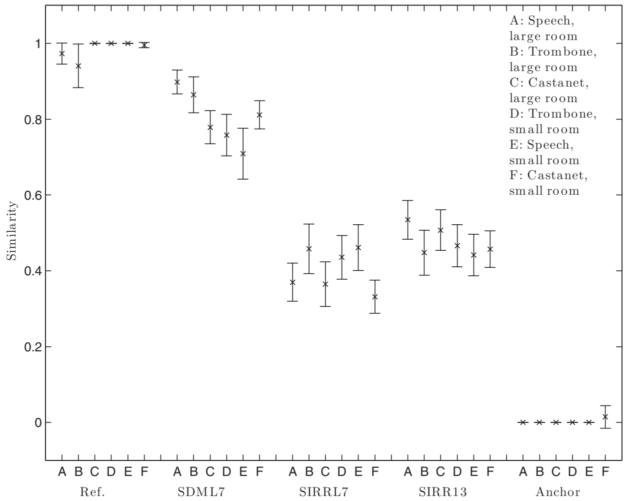


Fig. 7. Rated similarity of different spatial encoding method for each sample individually. The results are presented with mean and 95% confidence intervals.

respectively. For this reason, the anchor and the reference are removed from the ANOVA examination and the statistical tests are run again. Indeed, when these two methods are removed, Levene's test shows that the variances can be assumed equal [F(2,609) = 0.73, $p > 0.05$], and the Anderson-Darling test statistic indicates that the residuals are normally distributed [$A^2* = 0.99, p < 0.05$]. This means that the ANOVA is suitable for the data.

The results of the ANOVA indicate that the only significant main effect is the spatial encoding method [Method, F(2,576) = 332.80, p < 0.001] and none of the interactions is found significant. In total, the model explains 56% of the variance. The main effect for spatial encoding method is very strong and it explains 49% of the variance, and thus the remaining 7% are non-significant effects.

The results are presented for the spatial encoding method, in Fig. 7 with means and 95% confidence intervals. As can be seen from Fig. 7, out of the spatial encoding methods, SDML7 is the most similar with the reference, SIRRL7 is the least similar, and SIRR13 is slightly more similar than SIRRL7. All the means are significantly different and the average values for the methods are: reference: 0.98, SDML7: 0.80, SIRR13: 0.48, SIRRL7: 0.40, and Anchor: 0.00. Thus, according to the listening test experiments, SDML7 is the most similar with the reference out of the tested encoding methods. In the best cases, in samples A and B (speech or trombone in large room), the results of the reference and SDML7 are not significantly different. In all the other samples, the SDML7 results are significantly different from the other methods and the reference. The furthest from the reference are the results for sample E (trombone in small room).

All the attributes from the interviews are listed in Table 5. They are grouped into two groups, spatial and timbral aspects. The interviews of the test subjects revealed that they most often used localization, spatial impression, muddiness, coloration, distance, and clarity as the attributes for rating the samples.

Table 5. The attributes that the subjects used for assessing the similarity according to the interviews. Most of the attributes are translated from Finnish to English.

Timbral aspects
Localization (×5), spatial impression (×4), the amount of reverberation (×4), distance (×3), spatial width, spaciousness, gating effect, perception of room size, reverberation time, artifacts in the reverberation, direct-to-reverberant ratio.
Spatial aspects
Coloration (×7), muddiness (×4), clarity (×4), low-frequency content (×2), pitch (×2), brightness (×2), tone, depth, frequency shift, differences in the direct sound, metallic reverb, artifacts at high frequencies, high-frequency content.

4 DISCUSSION

4.1 Advantages and Drawbacks of the SDM and Future Work

In rendering, SDM uses only one omni-directional impulse response. Therefore, the frequency response in the exact sweet spot is identical to the original one in the pressure microphone. That is, due to the direct use of the pressure microphone signal, no peaks or dips occur in the frequency response in the sweet spot. The spatial distribution of sound can be inaccurate, but as the direct sound and early reflections are accurately reproduced the perceived error is negligible. In addition, in SDM the diffuse part of the sound field is obtained automatically, whereas in SIRR the diffuse part of sound is reproduced with uncorrelated loudspeaker signals, which are not easy to implement.

SDM assumes wideband signals. That is, the room impulse responses should be measured with full band width. In the case of band limited room impulse responses, SDM will artificially increase the energy outside the frequency band, since each sample in the encoded version is presented by a Dirac-impulse.

In a real room impulse response, the energy in the high frequencies decreases as time progresses due to air absorption and surface absorptions. Thus, the frequency response of the late reverberation is a "low-pass" filtered version of the original response of the direct sound. As pointed out by the listening test subjects, SDM slightly increases the perceived brightness or high frequency content in the late part of the impulse response. The division of the pressure signal causes this drawback. An image-source represents each of the samples in the pressure signal and the image-source is a Dirac-impulse in time-domain, which is wide band in frequency domain. Since the late part of the impulse response does not have as much energy on the high frequencies as the early part this results in an increase in the perceived brightness of the reverberation.

The problem of increased brightness in the late reverberation can be overcome by equalizing the frequency response in a post-processing step. Another option is to analyze the locations of the image-sources in frequency domain. This way, each frequency would have a correct weighting to begin with. This requires additional research and is therefore left for future work. In addition, future work includes open source implementations of the SDM encoding for a general pressure microphone array, B-format microphone, and the decoding implementations for wave field synthesis and higher order Ambisonics. The problem of increased clarity may not be present in the other reproduction approaches.

In this paper the room acoustic simulation used ideal specular reflections, which is an inherent property of the applied image-source room simulation method. SDM should also be tested with diffuse reflections. However, the generation of the reference case for diffuse reflection is problematic, since for the reference case the direction, time of arrival, and pressure value for each time instant is required. This information is available in beam-tracer or ray-tracing methods. Unfortunately, these methods neglect the temporal spreading of the reflections and consider that diffuse reflections only introduce spatial spreading for the reflected sound. Moreover, the room acoustic simulation methods that aim to solve the wave equation, e.g., finite element method, boundary element method, and finite difference in time-domain, may generate the correct pressure values, but they do not produce directional information. The only method that produces all the necessary information and takes into account the temporal and spatial spreading is presented in [35], but it only applies for low-frequencies. The comparison for a reference case with diffuse reflections is currently not possible.

SIRR was implemented with the parameters given in the original paper [26]. It should be emphasized that the advances made in Directional Audio Coding could possibly improve quality of the SIRR. In informal listening, for example, the multi-rate implementation [36] was found to increase the overall quality in SIRR. Studies using SIRR with alternative processing approaches are currently not available in the literature.

5 CONCLUSIONS

This paper presented a spatial encoding method for spatial room impulse responses. The analysis of the method estimates the location in very small time windows at every discrete time sample, where the localization method depends on the applied microphone array and acoustic conditions. Each of the discrete time samples is therefore represented by an image-source. Thus, the analysis results in a set of image-sources. Then, depending on the spatial reproduction method, the samples are distributed to several reproduction channels to obtain individual impulse responses for all reproduction channels.

The main advantage of the method follows from the decomposition of the image-sources. Namely, the method can be applied to any arbitrary microphone array and the spatial reproduction method can be any of a variety of existing techniques. It should be emphasized that the method is not designed for a continuous signal, but for spatial room impulse responses, which can then be convolved with an anechoic signal. In this paper the applied microphone array was an open spherical microphone array with six microphones, with an additional seventh microphone in the geometric center of the array.

Listening test experiments showed that the presented method produces sound that is indistinguishable from a reference sound in the best case. In overall, the similarity of the sound samples encoded with the presented method were perceived to be closer than that of a state-of-the-art method in the same conditions.

6 ACKNOWLEDGMENT

This work was supported by ERC grant agreement no. [203636] and Academy of Finland agreement nos. [218238 and 140786]. The authors are grateful to Prof. Lauri Savioja, Prof. Ville Pulkki, and Mr. Mikko-Ville Laitinen for discussions. The authors would also like to thank the anonymous reviewers for their valuable comments, which helped to improve the quality of this paper. Dr. Alex Southern is thanked for providing the implementation for air absorption filter computations.

7 REFERENCES

[1] T. Lokki, H. Vertanen, A. Kuusinen, J. Pätynen, and S. Tervo “Concert Hall Acoustics Assessment with Individually Elicited Attributes,” *J. Acoust. Soc. Am.*, vol. 130, pp. 835–849 (Aug. 2011).

[2] T. Lokki, J. Pätynen, S. Tervo, S. Siltanen, and L. Savioja, “Engaging Concert Hall Acoustics Is Made Up of Temporal Envelope Preserving Reflections,” *J. Acoust. Soc. Am.*, vol. 129, pp. EL223–EL22 (Apr. 2011).

[3] J. Daniel, R. Nicol, and S. Moreau, “Further Investigations of High Order Ambisonics and Wavefield Synthesis for Holophonic Sound Imaging,” presented at the *114th Convention of the Audio Engineering Society*(2003 March), convention paper 5788.

[4] A. J Berkhout, D. De Vries, and P. Vogel, “Acoustic Control by Wave Field Synthesis,” *J. Acoust. Soc. Am.*, vol. 93, no. 5, pp. 2764–2778 (1993).

[5] M. M. Boone, E. N. G. Verheijen, and P. F. Van Tol, “Spatial Sound-Field Reproduction by Wave-Field Synthesis,” *J. Audio Eng. Soc.*, vol. 43, pp. 1003–1012 (1995 Dec.).

[6] D. Hammershøi and H. Møller, *Communication Acoustics*, chapter 9, “Binaural Technique—Basic Methods for Recording, Synthesis, and Reproduction”, pp. 223–254 (Springer-Verlag, New York, NY, USA, 2005).

[7] D. Schönstein and B. F. G. Katz, “Variability in Perceptual Evaluation of HRTFs,” *J. Audio Eng. Soc.*, vol. 60, pp. 783–793 (2012 Oct.).

[8] V. Pulkki, “Virtual Sound Source Positioning Using Vector Base Amplitude Panning,” *J. Audio Eng. Soc.*, vol. 45, pp. 456–466 (1997 Jun).

[9] F. Zotter and M. Frank, “All-Round Ambisonic Panning and Decoding,” *J. Audio Eng. Soc.*, vol. 60, pp. 807–820(2012 Oct.).

[10] J. Merimaa and V. Pulkki, “Spatial Impulse Response Rendering I: Analysis and Synthesis,” *J. Audio Eng. Soc.*, vol. 53, pp. 1115–1127 (2005 Dec.).

[11] J. B. Allen and D. A. Berkley, "Image Method for Efficiently Simulating Small-Room Acoustics," *J. Acoust. Soc. Am.*, vol. 65, no. 4, pp. 943–950 (1979).

[12] U. P. Svensson, R. I. Fred, and J. Vanderkooy, "An Analytic Secondary Source Model of Edge Diffraction Impulse Responses, *J. Acoust. Soc. Am.*, vol. 106, pp. 2331–2344 (1999).

[13] B. I. Dalenbäck, M. Kleiner, and P. Svensson, "A Macroscopic View of Diffuse Reflection, *J. Audio Eng. Soc.*, vol. 42, pp. 793–807 (1994 Oct.).

[14] J.-M. Jot, L. Cerveau, and O. Warusfel, "Analysis and Synthesis of Room Reverberation Based on a Statistical Time-Frequency Model, presented at the *103rd Convention of the Audio Engineering Society, Convention*(1997 Sept.), convention paper 4629.

[15] C. Zhang, Z. Zhang, and D. Florêncio, "Maximum Likelihood Sound Source Localization for Multiple Directional Microphones,"*IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, pp. 125–128(2007).

[16] D. Levin, E. A. P. Habets, and S. Gannot, "Maximum Likelihood Estimation of Direction of Arrival Using an Acoustic Vector-Sensor," *J. Acoust. Soc. Am.*, vol. 131, no. 2, pp. 1240–1248 (2012).

[17] S. Tervo, "Direction Estimation Based on Sound Intensity Vectors,"*European Signal Processing Conference, Glasgow, Scotland, August 24-28*, pp. 700–704 (2009).

[18] S. Tervo, J. Pätynen, and T. Lokki, "Acoustic Reflection Localization from Room Impulse Responses," *Acta Acustica united with Acustica*, vol. 98, pp. 418–440 (2012).

[19] A. Host-Madsen, "On the Existence of Efficient Estimators," *IEEE Trans. Signal Processing*, vol. 48, no. 11, pp. 3028–3031 (2000).

[20] C. Knapp and G. Carter, "The Generalized Correlation Method for Estimation of Time Delay," *IEEE Trans. Acoust., Speech and Signal Proc.*, vol. 24, no. 4, pp. 320–327 (1976).

[21] L. Zhang and X. Wu, "On Cross Correlation Based-Discrete Time Delay Estimation,"*IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 4, pp. 981–984 (2005).

[22] T. Pirinen, *Confidence Scoring of Time Delay Based Direction of Arrival Estimates and a Generalization to Difference Quantities*, Ph.D. thesis, Ph.D. thesis, Tampere University of Technology, 2009. Publication; 854. Publication; 854.

[23] H. Kutruff, *Room acoustics, 4th Ed.*(Spon Press, NY, NY, USA, 2000).

[24] S. Tervo, T. Lokki, and L. Savioja, "Maximum Likelihood Estimation of Loudspeaker Locations from Room Impulse Responses," *J. Audio Eng. Soc.*, vol. 59, pp. 845–857 (2011 Nov.).

[25] A. Brutti, M. Omologo, and P. Svaizer, "Multiple Source Localization Based on Acoustic Map Deemphasis," *EURASIP J. Audio, Speech, and Music Processing*, 2010, 2010, paper 147495.

[26] V. Pulkki and J. Merimaa, "Spatial Impulse Response Rendering II: Reproduction of Diffuse Sound and Listening Tests," *J. Audio Eng. Soc.*, vol. 54, pp. 3–20 (2006 Jan./Feb.).

[27] J. Pätynen, S. Tervo, and T. Lokki, "Analysis of Concert Hall Acoustics via Visualizations of Time-Frequency and Spatiotemporal Responses," *J. Acoust. Soc. Am.*, vol. 133, no. 17 (January 2013).

[28] J. S. Bradley, H. Sato, M. Picard, et al., "On the Importance of Early Reflections for Speech in Rooms," *J. Acoust. Soc. Am.*, vol. 113, no. 6, pp. 3233–3244 (2003).

[29] Geneva International Telecommunication Union. ITU-R BS.1116-1: Methods for the Subjective Assessment of Small Impairments in Audio Systems including Multichannel Sound Systems, 1997.

[30] H. E. Bass, H.-J. Bauer, and L. B. Evans, "Atmospheric Absorption of Sound: Analytical Expressions," *J. Acoust. Soc. Am.*, vol. 52, no. 3B, pp. 821–825 (1972).

[31] T. Lokki, J. Pätynen, A. Kuusinen, and S. Tervo, "Disentangling Preference Ratings of Concert Hall Acoustics Using Subjective Sensory Profiles," *J. Acoust. Soc. Am.*, vol. 132, pp. 3148–3161 (Nov. 2012).

[32] J. Vilkamo, T. Lokki, and V. Pulkki, "Directional Audio Coding: Virtual Microphone-Based Synthesis and Subjective Evaluation," *J. Audio Eng. Soc.*, vol. 57, pp. 709 (2009 Sept.).

[33] B. B. Schultz, "Levene's Test for Relative Variation," *Systematic Biology*, vol. 34, no. 4, pp. 449–456 (1985).

[34] M. A. Stephens, "EDF Statistics for Goodness of Fit and Some Comparisons," *J. Am. Statistical Assoc.*, vol. 69, no. 347, pp. 730–737 (1974 Sept.).

[35] S. Siltanen, T. Lokki, S. Tervo, and L. Savioja, "Modeling Incoherent Reflections from Rough Room Surfaces with Image Sources," *J. Acoust. Soc. Am.*, vol. 132, pp.4604–4614 (June 2012).

[36] T. Pihlajamäki and V. Pulkki, "Low-Delay Directional Audio Coding for Real-Time Human-Computer Interaction," presented at the *130th Convention of the Audio Engineering Society* (2011 May), convention paper 8413.

THE AUTHORS

Sakari Tervo

Jukka Pätynen

Antti Kuusinen

Tapio Lokki

Dr. Sakari Tervo is a post-doctoral researcher in the Department of Media Technology, Aalto University School of Science from where he also received a D.Sc. degree in acoustics in January 2012. The topic of his research is on the objective room acoustic measures.

Previously he has been working in the Department of Signal Processing, Tampere University of Technology from where he also graduated as a M.Sc. majoring in audio signal processing in 2006. He has visited the Digital Signal Processing Group of Philips Research, Eindhoven, The Netherlands, in 2007 and the Department of Electronics of the University of York, United Kingdom, in 2010.

●

Dr. Jukka Pätynen was born in 1981 in Espoo, Finland. He received M.Sc. and D.Sc. (Tech.) degrees from the Helsinki University of Technology, Finland, in 2007, and Aalto University, Finland, in 2011, respectively. He is currently working as a post-doctoral researcher in the Department of Media Technology, Aalto University. His research activities include room acoustics, musical acoustics, and signal processing.

●

Antti Kuusinen received an M.Sc. degree in March 2012. He is currently working as a doctoral candidate at the Department of Media Technology at Aalto University School of Science under supervision of professor Tapio Lokki. In his doctoral research he focuses on the perceptual characteristics of concert hall acoustics. This research also includes development of descriptive vocabulary for sound, music, and room acoustics as well as elaboration of listening test methodology and statistical data analysis.

●

Dr. Tapio Lokki was born in Helsinki, Finland, in 1971. He has studied acoustics, audio signal processing, and computer science at the Helsinki University of Technology (TKK) and received an M.Sc. degree in electrical engineering in 1997 and a D.Sc. (Tech.) degree in computer science and engineering in 2002.

At present Dr. Lokki is an Associate Professor (tenured) with the Department of Media Technology at Aalto University. Prof. Lokki leads the virtual acoustics team jointly with Prof. Lauri Savioja. The research aims to create novel objective and subjective ways to evaluate concert hall acoustics. In addition, the team develops physically-based room acoustics modeling methods to obtain authentic auralization. Furthermore, the team studies augmented reality audio. The team is funded by the Academy of Finland and by Prof. Lokki's Starting Grant from the European Research Council (ERC).