OPEN ACCESS Freely available online

Synthesis of Spatially Extended Virtual Sources with Time-Frequency Decomposition of Mono Signals

TAPANI PIHLAJAMÄKI, *AES Student Member*, **OLLI SANTALA,** *AES Student Member*, **AND**
(tapani.pihlajamaki@aalto.fi) (olli.santala@aalto.fi)

VILLE PULKKI, *AES Fellow*
(ville.pulkki@aalto.fi)

Aalto University, Department of Signal Processing and Acoustics, Helsinki, Finland

Synthesis of volumetric virtual sources is a useful technique for auditory displays and virtual worlds. This task can be simplified into synthesis of perceived spatial extent. Previous research in virtual-world Directional Audio Coding has shown that spatial extent can be synthesized with monophonic sources by applying a time-frequency-space decomposition, i.e., randomly distributing time-frequency bins of the source signal. However, although this technique often achieved perception of spatial extent, it was not guaranteed and the timbre could degrade. In this article this technique is revisited in detail and the effect of different parameters is examined to ultimately achieve optimal quality and perception in all situations. The results of a series of informal and formal experiments are presented here, and they suggest that the revised method is viable in many cases. There is some dependency on the signal content that requires proper tuning of parameters. Furthermore, it is shown that different distribution widths can be produced with the method as well. From a psychoacoustical perspective, it is interesting that distributed narrow frequency bands form a spatially extended auditory event with no apparent directional focus.

0 INTRODUCTION

Auditory displays are used to synthetically present a perceivable sound scene to a user. They have applications in scientific research (e.g., in psychophysics), user interfaces, as well as in recreation like video games and virtual realities. There are several different synthesis techniques to render the sound scene for the user, and volumetric virtual source synthesis is one of them.

Volumetric virtual sources are virtual sources that have a spatial volume, in contrast to point-like virtual sources. Traditionally, point-like virtual sources have been used to synthesize virtual sound scenes, but volumetric virtual sources offer flexibility, realism, and creative potential for sound scene design. From the user point of view, the perception of a volumetric virtual source can be simplified into a perception of spatial extent because the ability to accurately discern distances of unfamiliar sounds is not good [1].

In this paper a method for synthesizing the perceived spatial extent for monophonic input is studied and revised. The main objective of the study is to find out why the method works well in some situations and how to modify it to perform well in all situations. In addition, the presented method has been originally used in the context of Directional Audio Coding research, but now the method is formulated to be independent of the reproduction method.

This paper is organized as follows. First, the background is discussed by describing similar existing techniques, perception of spatial extent, and the relation of perceived spatial extent to the apparent source width. This is followed by a description of the method with a discussion of its parameters and their effects based on informal experiments. Next, two formal listening experiments are described and their results are presented. Finally, the paper ends with a discussion of the implications of the results, aspects of the method, and psychoacoustical questions raised by the study.

1 BACKGROUND

1.1 Research on Spatial Extent Synthesis

There has been some previous research in synthesizing or modifying the perceived spatial extent of virtual sources. The method revised in this paper was originally presented in a conference article by Pulkki et al. [2]. It introduced the method in the context of virtual world Directional

Audio Coding (VW-DirAC). The presented method divided a monophonic input signal into frequency bands with equivalent rectangular bandwidth (ERB) filters and then randomly placed these frequency bands into different directions in azimuth angle using an angular distance limiting rule (i.e., frequency bands cannot be too far from each other). The result was reproduced using monophonic Directional Audio Coding (DirAC) reproduction with directional metadata provided by the assigned directions.

This work was further extended in the article by Laitinen et al. [3]. In this version, a synthetic B-format signal was created from the assigned directions and processed with B-format DirAC reproduction. The direction assignment was fine-tuned by relaxing the limiting of angular distance and creating an algorithm that ensured the filling of the desired extent. In addition, the article included formal listening tests that confirmed the plausibility of the method for synthesizing spatial extent with selected signal content.

Verron et al. [4] have performed research on an immersive environmental sound synthesis. Their aim was to completely synthesize different environmental sounds, i.e., fire, wind, and rain, and also synthesize the extent of these sounds. In the end, their approach achieves spatial extent by synthesizing multiple incoherent versions of the sound and mixing between them. Although their approach is quite different compared to the methods used in this study, it achieves similar results.

Another approach to synthesizing perceived spatial extent is to generate multiple incoherent point sources by decorrelating the original source signal and then placing these incoherent sources to different spatial locations (e.g., loudspeakers). Potard and Burnett [5] studied this and demonstrated that it is a viable method for controlling the perceived spatial extent and can also be used to control the extent and coherence of a virtual source on a sub-band basis. This results in different frequency bands having different extents and positions. They noted that this effect is easily perceived after a substantial amount of training.

A recent proposal in virtual source extent control is by Zotter et al. [6]. They formulate their earlier research in stereo phantom-source widening into a frequency-varying Ambisonics encoder. This can be thought as a spatial extent synthesizer, comparable in applicability to the one presented in this paper. Although they only presented experiment results for a two-dimensional spatial extents, they did propose a method for three-dimensional spatial extents.

Another recent proposal is by Pestana and Reiss [7]. They present a method that applies short-time Fourier transform similarly as in the present paper to decompose a sound stream in the time-frequency-space domain. They use this method for automatic adaptive spectral unmasking in stereo mixing. This leads to different methodology compared to the present article but their work and results can be considered complementary to the present paper.

1.2 Perception of Spatially Extended Sources

The aim in the presented method and the listening experiments of this article is to study the construction of spatially extended virtual source constellations that would result in the perception of extended auditory events. Thus, in this section, the perception of spatially wide, distributed, or extended sound source constellations is discussed. Different attributes that have been found to affect the perception and contribute to the widening or spreading of the perceived auditory event are presented.

The basis of human sound localization lies in the interaural cues, namely time and level differences (ITD and ILD, respectively), caused by the fact that sound coming from the side arrives earlier and at a higher sound pressure level to the ear that is closer to the sound. In addition, monaural cues caused by the reflections of the external ear and torso aid in localization.

Relevant to localization, especially in reverberant environments, is the precedence effect [8], a group of phenomena related to the human ability to localize sound based on the first-arriving component in the presence of reflections and reverberation. In summing localization, concurrently presented coherent signals form a virtual sound source between the original sources [9]. With lower coherence, the signals may be perceived as separate auditory objects, or they may form a spatially extended auditory event. Such effects were studied by Blauert and Lindemann [10] in a headphone experiment with altering interaural cross-correlation (IACC) values. With an IACC of one, i.e., equal headphone signals, one auditory event was perceived in the center of the head, while with an IACC of zero, one auditory event was perceived near each ear. Values between one and zero caused a perception of a spatially spread auditory event of varying extent inside the head.

When the spatial distribution of the sound event is complex, perception of the scenario is more challenging than with only one or two sound sources. There are several attributes that affect the perceived width or extent of a sound source ensemble—it depends, e.g., on the spectral content of the signals, temporal duration, and the number of sound sources presented simultaneously [9, 11–13]. In headphone listening, the perceived width has been shown to increase as the sound pressure level or temporal duration of the signal increases [14].

Next, a number of studies closely related to the listening experiments of this article are discussed in more detail. In an experiment by Hirvonen and Pulkki [15], wideband noise was divided into narrow bands and presented to participants from nine distributed loudspeakers in the frontal horizontal plane with a span of 45°. The frequency bands were always spatially placed in ascending or descending order, while the location of the starting point—the lowest frequency band—was changed from one case to another. The participants indicated the perceived center of gravity as well as all the loudspeakers they perceived as emitting sound. The results indicated that the perceived center of gravity tended to be near the loudspeakers from which the highest and lowest frequencies were emitted. In other words, either those frequency bands or the irregularity in frequency drew attention. The perceived width was never more than half of the width of the loudspeaker ensemble, suggesting that some frequency bands were fused together spatially.

Further proof for these findings was found in another study by Hirvonen and Pulkki [16], where eleven loudspeakers were distributed at $\pm 45^\circ$ and narrow frequency bands were routed to varying loudspeakers. Again, the lowest and highest frequency bands had a more significant effect on the results than the middle bands. Especially interesting in the scope of the present work was that the test cases where the adjacent frequency bands were not in neighboring loudspeakers were perceived as being slightly wider than those where the frequency bands were spatially placed in ascending order.

Santala and Pulkki [13] studied the perception of spatially distributed sound sources using pink noise and thirteen equidistant loudspeakers placed every 15° on the frontal horizontal plane, forming a distribution of $\pm 90^\circ$. Different distribution widths as well as cases with gaps in the distributions were included. The participants were allowed to rotate their heads but were not allowed to move otherwise. The results indicated that accurately perceiving the distribution of the sound sources was challenging when more than three loudspeakers were simultaneously emitting sound. The width of the distribution could be perceived quite accurately, but the perceived width was slightly narrower than the actual width. In general, fewer loudspeakers were indicated as emitting sound than were actually emitting sound.

Hiyama et al. [17] used a setup with 24 loudspeakers on the horizontal plane around the listener to investigate the number of loudspeakers needed to reproduce a diffuse spatial impression. Different loudspeaker combinations were compared to a reference with all 24 loudspeakers emitting sound. With an evenly spaced, surrounding layout with uncorrelated white noise signals, six loudspeakers were enough to reproduce a similar perception as with 24 loudspeakers. With bandpass noise, it was found that when the layout was optimized, even four loudspeakers caused a perception close to that of the reference. Matching results were found in the second experiment of Santala and Pulkki [13]—with broadband noise it is easier to perceive differences in spatial distribution than with narrowband noise.

1.3 Measures of Spatial Extent and Envelopment

There are a number of different terms and attributes that are related to the spatial properties of sound in room acoustics. Here, the terminology selected for the present study is briefly justified. Generally, the perceived width of a frontal sound source group may be addressed with terms apparent/auditory source width and ensemble width, whereas the overall perception of the sound around the listener may be described by listener envelopment, spaciousness, spatial impression, and surroundedness [18, 19]. These terms are typically used when discussing the acoustics of concert halls as well as when the connections between the acoustical measures of rooms and these perceptual terms are studied [20, 21]. Since the listening experiments in this article were conducted in an anechoic chamber, all the aspects of the measures related to room acoustics are not directly applicable in the cases where the scenario does not include reflections.

In this article the area to be investigated has been chosen to be described as perceived spatial extent. This term is the most suitable one for explaining the attributes of the auditory event that the method aims to reproduce. Auditory source width and other terms related to the width of a scenario are prominently used to describe sounds in the field of view of the listener and do not take into account the sounds that are around the listener. On the other hand, terms related to sounds surrounding the listener tend to concentrate on the overall feeling of sound everywhere around the listener. Perceived spatial extent includes both sides of the aforementioned terms in the sense that the sound may be surrounding or it may have a specific width and, additionally, there may be perceived centers of gravity or specific directions of emitted sound. Similar reasoning for selecting perceived spatial extent as the term to be used can be found in [22]. On a side note, the focus in this article is not on depth, whether it be ensemble or environment depth, nor on the perception of distance, leaving these attributes outside of the scope of this discussion.

2 METHOD

As mentioned in Sec. 1.1, the method used in this study is based on the original concept presented in [2]. The intention of this study is to explore why the proposed method works and how it could be further developed into a practical tool for sound design. Moreover, the original method did not always work as intended and could sometimes compromise timbral quality. For the purpose of improving the method, a multitude of different parameters are discussed in this section in addition to explaining the basic method. Variations of these parameters were tested with informal listening by the authors to judge their effects on different signal types. In addition, a few informal listening experiments were performed with a small number of listeners to find more information about personal preferences. Any comments about the effects of parameter variations in this section can be assumed to be derived from these informal experiments and, unless explicitly defined, these experiments were performed on a horizontally surrounding reproduction system. The most important effects have been selected for formal evaluation in Secs. 3 and 4.

2.1 Description of the Method

The algorithm itself is very simple, as is illustrated in Fig. 1. First, a monophonic input signal is inserted into the system and a time-frequency transform is performed on it. The resulting time-frequency domain signal is then given to the following spatial distribution algorithm:

1. Start from the lowest frequency band.
2. Select (randomly or deterministically) the direction for the current frequency band.
3. Create a bandpass filter for the current frequency band.

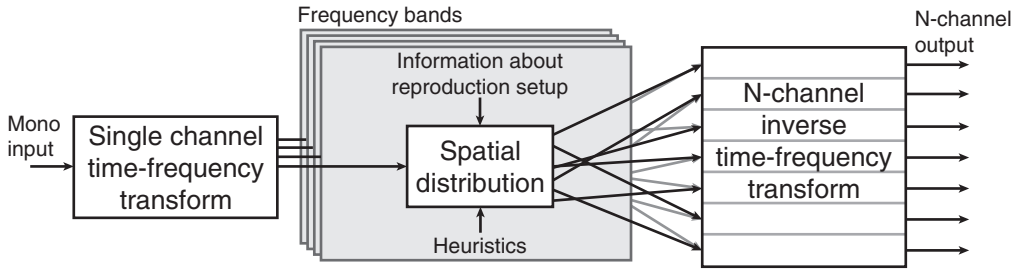

Fig. 1. Block diagram of the extent synthesis algorithm.

4. Sum the current passband filter into the final filter corresponding to the selected direction.
5. Select the next frequency band and loop from step 2 until all frequency bands have been processed.

The next step is to filter the monophonic signal with the produced final filters. This generates a multichannel signal with (in theory) no channel having common frequency content. The resulting multichannel signal can then be inverse transformed to produce the final output.

Although the algorithm is simple, it contains several parameters that can be adjusted. First, the time-frequency transform and its properties can affect the result. Second, the spatial distribution method can be designed in various ways. The assignment of directions can be random or deterministic, and these properties can be further inspected. In addition, there is the question of whether or not the assignment of directions should change through time. Third, the performance of the method can be signal-dependent and it is possible to preprocess the signal with (for example) a decorrelation algorithm to change its properties.

2.2 The Aim of This Study

The presented method generates a specific type of output signal. When the parameters are selected properly, the result is an evenly surrounding and enveloping auditory event that, in essence, has the sound coming from everywhere and nowhere. Furthermore, the sound does not significantly change with the head movement of the listeners. Although this type of perception can be disturbing when no direct sound is present, it was deemed by the authors that this property is the most desired one in this study. Thus, the aim is to create this effect with as many signal types as possible by adjusting the parameters. From the user point of view, the ideal solution would be a premade black box that generates the desired effect with any signal content without any adjustments to the parameters. The user can then freely combine it with direct sound and reverberation.

Based on these objectives, the parameters and their adjustments are discussed in the next sections.

2.3 Parameters of the Time-Frequency Transform

There are various choices on how to perform the time-frequency transform for the extent synthesis method. In essence, anything should work—short-time Fourier transform, wavelet transform, Gammatone filter bank, linear filter bank—but there are differences in the resulting output. For its simplicity and efficiency, short-time Fourier transform (STFT) was chosen as the main method in this work. Additionally, filter banks with linear and equivalent rectangular bandwidth (ERB) [23] spacing were tested.

2.3.1 Short-Time Fourier Transform

STFT offers several clear parameters that can be varied: time window length, overlap amount, and window function. It is possible that all these parameters can affect the final quality of output, but the time window length, or frequency resolution, has the most direct effect on the algorithm. This is because it directly affects the number of frequency bands that can be distributed in the distribution algorithm. Thus, it is a prime candidate for experimenting.

Experimentation was done by increasing the time window size in powers of two from 64 to 8192 samples. This suggested several effects, the most interesting being that increasing the resolution seems to make the perceived spatial distribution more even and surrounding while maintaining a perceptually pleasant timbre. However, a window size larger than 1024 samples does not seem to improve the spatial quality and can also incorporate “metallic” timbral artifacts into the sound. Another effect is the time smearing created by long filters used in this method. This is perceived as inaccurate and noisy onsets and can be disturbing already with a 512 sample window with signals containing impulsive sounds. These preliminary results suggest that time window sizes of 512 samples and 1024 samples should be evaluated formally, as they seem to be the best compromises between timbral and spatial quality. This evaluation will be presented in Sec. 3.

It has been found out in previous research that applying a multi-resolution approach with STFT can be beneficial for quality in spatial sound reproduction [24]. This makes the multi-resolution STFT interesting for the proposed algorithm as it follows better the human time-frequency resolution with a longer time window for low frequencies and a shorter time window for high frequencies. The informal experiments, however, suggest otherwise. Using multiple concurrent resolutions does not seem to give any additional benefits. On the contrary, the disadvantages (timbral artifacts, uneven spatial distribution) seem to manifest themselves. Thus, it was deemed that multi-resolution STFT is not worthy of further investigation in this context.

2.3.2 Filter Banks

STFT can be thought of as a linear filter bank with equal bandwidth and a specific order for each filter. These parameters vary based on the time window length. However, with normal linear FIR filter banks, it is possible to control these parameters more freely. Thus, experiments were performed with different linearly spaced and ERB-spaced filter banks. In this case three parameters were used: frequency scale, number of frequency bands, and filter order. This is interesting to explore as the original method successfully applied ERB filter banks in spatial extent synthesis [2], although unreliably.

Informal experiments suggest that for the aim of this study—even and surrounding output—ERB spacing is better. With the same number of frequency bands, ERB spacing seems to generate a perceptually more even distribution than linear spacing. As for the number of filter bands, increasing the number seems to create a better quality output until a certain limit. Nevertheless, it seems that several hundreds of frequency bands are still required for a result of perceptually good quality. The effect of filter order was not studied extensively, but short experiments suggested that increasing the filter order (and thus the separation between frequency bands) has a positive effect on quality, although time smearing also increases.

Overall, it seems that filter banks can produce quality that is equal or possibly even better than with the STFT approach. However, the disadvantage is in efficiency, as implementing filter banks in the time domain is generally inefficient; and with STFT implementation, one might as well directly use the produced STFT frequency bands.

2.4 Distributing Time-Frequency Content in Space

Spatial distribution of the time-frequency tiles (i.e., the frequency bands in each time window) can be performed in many ways. The chosen method affects the output quality directly and thus should be carefully selected. Furthermore, there is the question whether the distribution should change through time or not. This section discusses the possibilities in depth. All of the methods presented here assume that the time-frequency tiles are distributed based on the spatial direction on a circle surrounding the listener's position, i.e., there are no signal level differences caused by the algorithm between different tiles.

2.4.1 Distribution Methods

The distribution method can be random or deterministic. A trivial solution is to use a uniform random distribution for spatial locations of the time-frequency tiles. In this case, every tile gets a completely random direction in the output. If there are enough tiles to be distributed, this method is usually quite even. However, a random sample is indeed random and the result can randomly be perceptually very good (i.e., even and surrounding with pleasant timbre) or perceptually very bad (i.e., point-like with timbral artifacts and inside-the-head perception). This result is similar to the original method in [2].

Informal experiments with uniform random distribution suggested that in many perceptually good cases, the neighboring frequency bands were distributed far away from each other. This suggests a modification to the random distribution algorithm. A constraint was applied to the uniform distribution algorithm so that each frequency band has to be located at least 90 degrees apart from the previous one. The resulting distribution is still uniform although a bit more predictable in performance. Again, informal experiments seem to support this theory. This modified random method seems to be more reliable and seems to almost always produce perceptually good results.

Another approach is to use a deterministic distribution. The advantage of a deterministic method is that it is predictable and always produces the same distribution, thus producing the same output. Hirvonen and Pulkki [15] performed experiments by distributing neighboring frequency bands in an ascending order through spatial angle. This seemingly does result in perceptual width for the virtual source, but the perception is not even, the timbre is unpleasant, and there seems to be severe perceivable "phasing" artifacts when the listener turns their head. Thus, this method cannot be suggested for any actual use. Nevertheless, there are deterministic methods that are suitable for this purpose. Low-discrepancy sequences are number sequences that resemble the uniform random distribution but are deterministic and "more uniform." They are generated with specific algorithms and are usually used in statistic trials when representative sampling is desired with a low number of samples. Similarly, they should be a very good candidate for an even distribution.

In this study the Halton sequence [25] was selected. This sequence produces fractional numbers between 0 and 1 by dividing the number axis by multiples of the base value. Changing the base of the sequence produces different sequences. Thus, this sequence can be used for two- or three-dimensional cases, as desired. However, it was noted that this sequence does not work well for a circular distribution (e.g., angles on a surrounding circle) because around the wrap point there are very often close distance neighboring points. However, this problem can be avoided by using two different number sequences to create two-dimensional points and then converting from vectors to angles.

In this study the Halton sequence is applied so that base 2 and base 3 Halton sequences were generated and 52 first values were skipped. The skip value was selected by studying the energy distribution of different signal types with different window sizes and skip values. This value was then verified by informal listening. Nevertheless, other values should be perfectly suitable.

After creating the sequence, pairs of values are selected from the two sequences and scaled to fill the area from –1 to 1. These value pairs form 2D-points that can be assigned to frequency bands starting from the lowest. If a 2D-point would be outside the unit circle, it is discarded. Once all frequency bands have a spatial location, the assigned points are converted to angles, rounded to discrete loudspeakers, and used as the distribution for the time-frequency tiles. Fig. 2 shows this process and the differ-

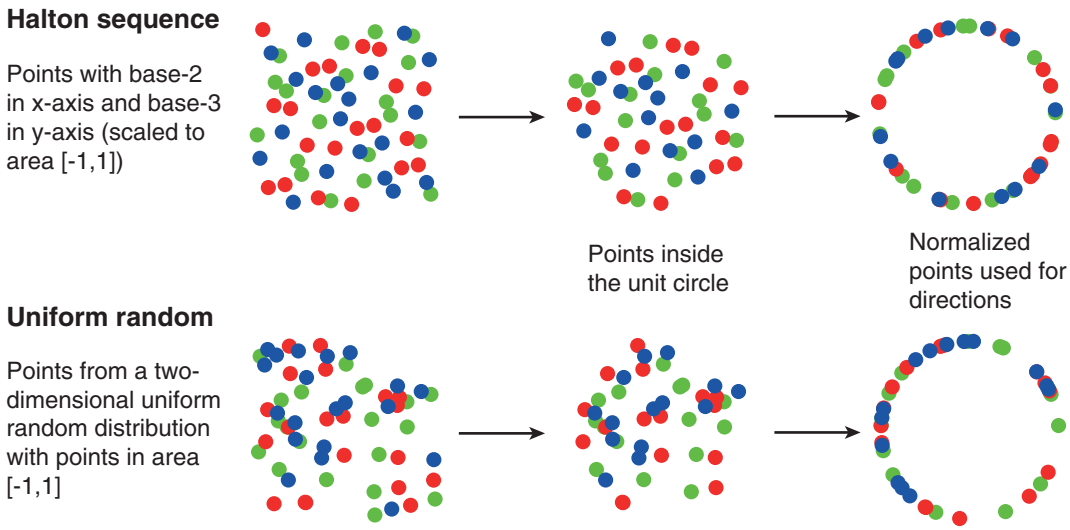

Fig. 2. Example of a difference between first 60 points generated from the Halton sequence and a uniform random distribution when they are applied in the algorithm. Color shows consecutive 20 points from the set. Note that the Halton sequence is more uniform overall and has consecutive points distributed all around the areas.

ence between the Halton sequence and a uniform random distribution.

Informal experiments suggested that the Halton sequence is very suitable for the task. The result seems to be generally even and surrounding. Furthermore, this method seems to perform better with shorter time windows than the random distribution method. However, even this method occasionally seems to produce some perceptual concentrations, depending on the signal content. Nevertheless, these can be usually avoided, if necessary, with a different skip value. Thus, this method is the best candidate for a stable distribution of time-frequency tiles.

2.4.2 Time Variant Distribution

The methods in the previous section assume that the distribution is static, i.e., after the initial setup, the distribution does not change. However, the distribution can be time variant. The advantage of this should be a more even perceptual distribution of frequency content with fewer or no perceivable separate frequency components. On the other hand, changing the distribution directly leads to transient artifacts caused by the sudden change in the filter coefficients.

Several informal experiments were performed using different schemes for time-variant distribution. They suggested that the best results could be produced with a suitably slow change of directions and changing only a few frequency bands at a time. However, a single informal blind experiment also suggested that the static distribution is usually preferred in direct pair-wise preference comparison to a time-variant one, and thus this method is not included in the formal experiments.

2.4.3 Non-Surrounding Spatial Extents

Often, a completely surrounding spatial extent is not desired. Instead, the sound source should have a clear perceivable extent, but it should fill only a certain portion of the perceivable space. Although the presented method is here mainly discussed in a completely surrounding context, it can be simply defined for other spatial extents. This is done by "cutting" the surrounding distribution and then mapping the "strip" into a smaller area or volume. For example, with a circular distribution, this is done with angles by using, for example, the equation

$$\theta_\alpha(k) = \theta_{360}(k)\frac{\alpha}{360}. \quad (1)$$

Here, α is the desired total spatial extent in degrees, $\theta_{360}(k)$ are the original surrounding distribution angles, and $\theta_\alpha(k)$ are the resulting distribution angles. This method works with all presented distribution methods and can be extended to three-dimensional cases alike.

2.5 Signal Dependency

As has been already mentioned, the presented method is dependent on the source signal content. Some input signal types tend to produce very easily perceptually surrounding and even output signals that do not contain any disturbing artifacts. This can even happen with many different parameter combinations. On the other hand, it is very hard to create perceptually surrounding output with some other input signal types while preserving the timbral quality.

To determine the effect of the signal content, several different recorded sound samples were processed with the method. Table 1 shows the different signal types that were tested during this study, their overall quality, strictness for selecting proper parameters, and possible specific artifacts that become apparent with bad parameters. It is evident from this table that many signal types can be processed with the method. But how does the signal content affect the output?

First of all, with most signals, the perceived timbre does change. This change is natural, as the perceived sound field

Table 1. Tested signal types, their overall quality, strictness to proper parameter selection, and specific artifacts that are apparent with bad parameters.

Signal content	Overall quality	Parameter strictness	Specific artifacts
Pink noise	Bad	Not possible	Comb-filtered sound
White noise	Bad	Not possible	Comb-filtered sound
Anechoic female speech	Good	Strict	Unnatural timbre, inside-the-head perception
Anechoic male speech	Good	Strict	Unnatural timbre, inside-the-head perception
Anechoic acoustic guitar	Excellent	Relaxed	Can be slightly implausible
Anechoic cello	Excellent	Relaxed	Can be slightly implausible
Anechoic conga drumming	Bad	Strict	Spread onsets, noisy timbre
Anechoic trombone	Excellent	Relaxed	None
Acoustic guitar in short reverberation	Excellent	Relaxed	None
Cello in short reverberation	Excellent	Relaxed	None
Conga drumming in short reverberation	Bad	Strict	Spread onsets, noisy timbre
Flying chopper	Very good	Relaxed	Above-the-head perception
Seashore	Very good	Average	Comb-filtered sound, unnatural spatial perception
Sailing boat at sea	Good	Average	Comb-filtered sound

becomes surrounding instead of coming mainly from a single direction. However, with some signals, this change is also unpleasant. For example, even in the best case, the pink noise does not sound pleasant. On the other hand, with suitable signals, it seems that the resulting sound can be, as a whole, perceived as better than the original signal as the spatial extent complements the timbral changes.

Second, depending on the time structure of the signal, the time smearing effect caused by high-order filters can be audible. This is especially prominent with sparse, impulse-like signals, e.g., conga drumming. Third, with some signals (e.g., speech), it is questionable how plausible the perceived sound can be. How does one imagine a single anechoic instrument or speech as surrounding? Thus, it is evident that some signals generally should not be processed with the presented method unless the side effects are especially desired.

As a general rule, it seems that signals with no sudden or impulsive events are most suitable for the method. Reverberation seems to make the processed signal more plausible. This is most likely due to the fact that spatially extended sound often contains reverberation. Furthermore, wide frequency content is preferable but not mandatory.

2.6 Decorrelation Pre-Processing and Multiple Distributions

Decorrelation, in the context of this paper, is a method for creating perceptually mutually incoherent versions from a single input signal. As was mentioned in Sec. 1.1, this method can be used to create spatially extended sources at a relatively high computational cost. However, decorrelating a signal affects its time-domain structure by scrambling the phase and removing possible phase alignment over frequency that might be audible [26]. This can be also thought of as adding a very short reverberation to the signal. As reverberation generally helps the presented method, it is possible that decorrelation could be used as a preprocessing method for signals to enable proper synthesis of spatial extent. Based on informal experiments, this seems to be true. It is easier (i.e., it works with multiple parameter combinations) to create spatial extent for a decorrelated signal. However, depending on the signal, the decorrelation filter can be audible in the output signal, which is not desirable.

Another method to apply decorrelation is to create multiple incoherent sources that are then spatially distributed in a defined pattern with the presented method. The simplest case is to create two sources. This method is shown in Fig. 3 and is further discussed as the mirror distribution. First, the input signal is transformed into the time-frequency domain, as with the normal method. Then, it is decorrelated with two different decorrelation filters. The next step is to create a normal spatial distribution for one of the signal paths. The information about the distribution is given to the other signal path where a mirror distribution is created. This mirroring is done through the point where the listener's avatar is located in the virtual space. In the most simple case, this is the center point of a circle surrounding the listener. Finally, both signal paths are multiplied with $\frac{1}{\sqrt{2}}$, added together, and the sum is inverse transformed for output. The decorrelation in this case is based on frequency-dependent random time delays [27] although other approaches should be equally suitable.

The resulting output signal appears to be in many cases perceptually surrounding and seems to have only few or no perceivable artifacts due to the decorrelation. Furthermore, this method seems to be more robust in relation to different parameter values. However, some "impact" and accuracy of onsets appears to be lost due to the removal of phase alignment. Nevertheless, this method will be studied formally in Secs. 3 and 4.

3 FORMAL EVALUATION—EXPERIMENT 1

A formal experiment was organized to evaluate the presented method. This experiment was divided into three parts, 1a, 1b, and 2, and this section will present the organization and results of 1a and 1b.

3.1 Experiment 1a—Evaluation of Spatial Perception

The aim of this experiment was to evaluate how the spatial qualities of the presented method are perceived by the

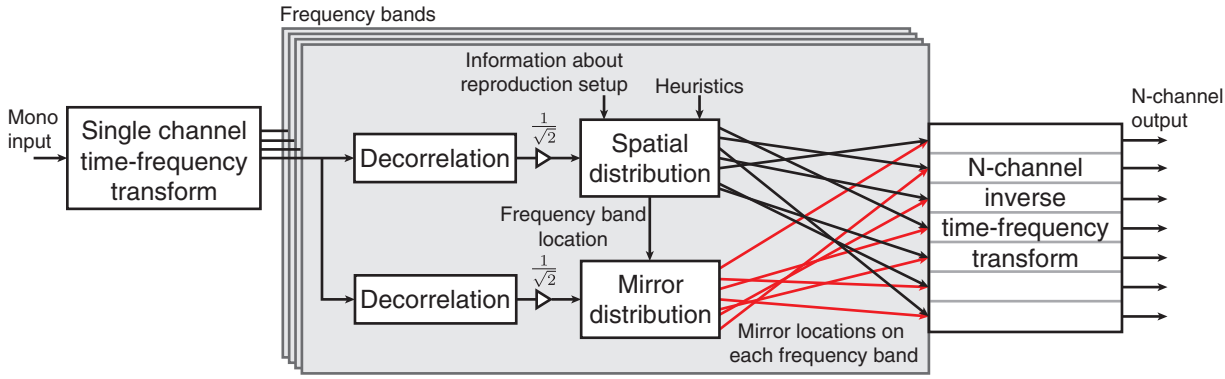

Fig. 3. Block diagram of the mirror distribution method.

experiment participants. In this case the reproduction was always intended to be completely surrounding and even. The task was to indicate for each direction with 30° spacing if there was any sound coming from that direction and if the sound was a clear specific part of the auditory event. Corresponding to this intention, the participants were instructed to use the terms "major," "minor," and "no sound" in the graphical user interface (GUI) (shown in Fig. 4), as follows:

Major: A specific part of the auditory event is localized in that direction.
Minor: There is some sound in the direction.
No sound: There is no sound in the direction.

This terminology was further explained in a separate training session that is described in Sec. 3.3. The aim was that "major" would be used when any direction could be perceived as a clear part separate from the overall sound, whereas "minor" would be used when any sound could be heard from that direction. By definition, "minor" is then implicitly included if "major" was marked. In this experiment the answer for an ideal spatial extent synthesizer would be all "minor."

In addition, it was possible to mark if the sound was perceived to come from inside or near the listener's head, or above or below the listener. Furthermore, it was possible to write comments on a paper during the experiment or give verbal comments to the organizer during breaks.

Three different parameters were selected for this experiment: time-frequency resolution in terms of STFT window length (256, 512, 1024, and 2048 samples), distribution method (Halton sequence and uniform random distribution), and processing technique (mirror distribution synthesis, unprocessed synthesis, and decorrelated synthesis; as mentioned in Sec. 2.6, decorrelation is based on random time delays as per Bouéri and Kyriakakis [27]). Four different program materials were chosen as representative candidates of different interesting sample types: anechoic cello music, cello music in short

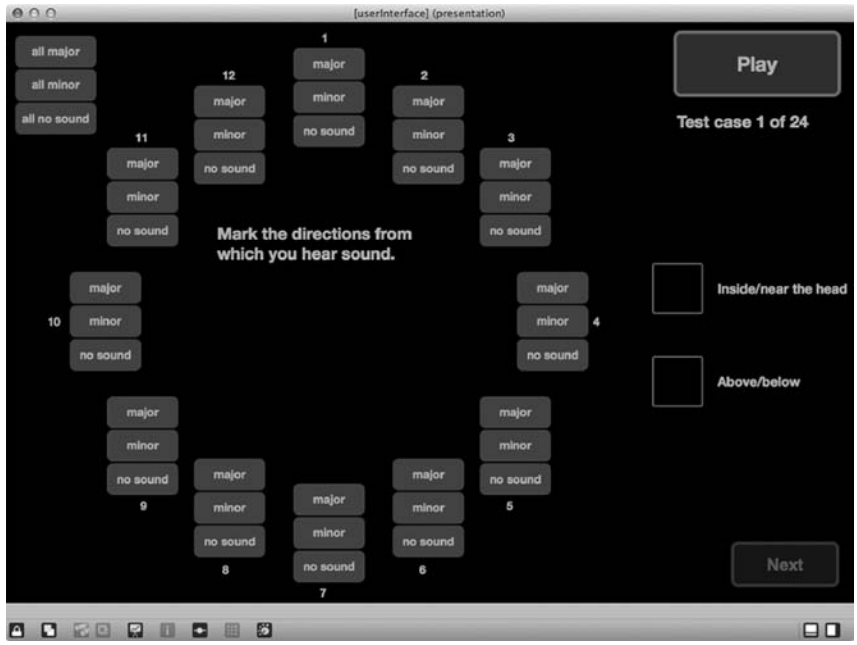

Fig. 4. User interface for experiment 1a.

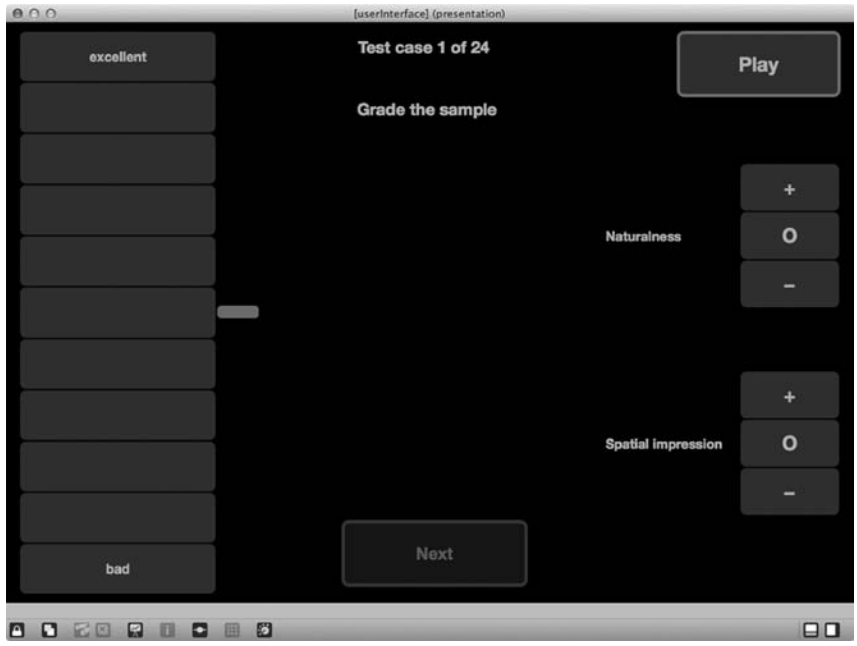

Fig. 5. User interface for experiment 1b.

reverberation, anechoic male speech, and recording of a seashore.

Each parameter combination was synthesized separately for each participant, i.e., every participant had a different rendering of the signal output. This was done to study the difference between random distribution and deterministic distribution. Each participant graded each parameter combination once.

3.2 Experiment 1b—Evaluation of Subjective Preference

The aim of this experiment was to evaluate the perceived quality of the signals processed with the presented method. This was organized as a preference test. The task of the participants was to indicate their preference of the sample on an 11-step scale from "excellent" to "bad" using the GUI (shown in Fig. 5). These anchor point words were defined for the participants as follows.

Excellent: In a context or situation where this sound is appropriate, you would prefer no other reproduction.
Bad: In general, you would not prefer to hear this sound in any situation.

It should be noted that the task was not to rate the performance of the method itself (i.e., how well synthesis of the spatial extent works in each case), but to rate the overall preference of the signals, thus including the personal opinion of the source signal.

In addition, the participants could mark how the "naturalness" and "spatial impression" affected their grading on a simple "positive," "neutral," "negative" scale. These two factors were found to be appropriate in a preliminary informal listening session where some of the final experiment samples were played for naive listeners without any other preparation than instruction to describe the perceived sound freely. Furthermore, participants could write comments on a paper and give verbal comments to the organizer.

The parameter combinations and program material were the same in this experiment as in experiment 1a. Additionally, the individual renderings of the signal output were the same in 1a and 1b for a single participant. Thus, each participant answered for both spatial impression and preference score based on the same signal content.

3.3 Experiment Organization

The experiments were organized in an anechoic chamber with an even-spaced, horizontal loudspeaker setup. The lights were turned off during the experiment, and the loudspeakers were marked with illuminated numbers from one to twelve to represent direction. The same numbers were shown in the GUI as well. Participants sat at the center of the room in the sweet spot of the loudspeaker setup. They were free to rotate with the chair and look in any direction during the experiments. However, they were asked to keep sitting normally and not to move towards the loudspeakers. In addition, for each test case, they were asked to rotate at least one full circle and advised to freely rotate while listening. Prior to each test case, there was an orientation sound from a random direction, towards which the participant was to face before starting the playback of the test case. A GUI on a tablet device was used for answering, and paper and pen were provided for additional comments.

There was a total of 24 participants in experiments 1a and 1b. Their age was between 24 and 37 years. All had previous experience in listening experiments or analytical listening in general. None reported any severe hearing disorders that would affect the experiment. The authors did not participate in the experiments.

Experiment 1 was organized in six one-hour sessions where the participants took one part of experiment 1a or two parts of experiment 1b. There was one break included

in each session. One of the participants had to do an additional session as he used more time than was expected in his first part of experiment 1b. Before each part, the participant listened once through all of the cases used in the part. This took approximately four minutes. One part of experiment 1a took on average 45 to 60 minutes, and one part of experiment 1b took on average 15 to 25 minutes. Each participant had to assess 24 cases in each part.

Experiments 1a and 1b were organized so that half of the participants first performed 1a in full and the other half started with 1b. Participants assessed one program material with different parameter combinations in one part. This was done to make the task easier for the participants, as the differences could be very small in some cases. Within experiments, a balanced latin square design was used to select the presentation order of program material for each participant. However, this design is not fully balanced, and thus it has to be taken into account in the statistical analysis. Nevertheless, this procedure simplified test organization.

Before both experiments 1a and 1b, a short training session was organized where seven differently processed pink noise signals were played and explained to the listeners to attune them to possible differences in the experiment cases. These signals were not processed with the presented method and the examples contained both single and multiple level differences or colorations. Before participating in experiment 1a, two of the presented examples also showcased the "major" and "minor" cases. In addition, this session acted as a screening to ensure that the participant could hear the differences.

3.4 Statistical Tests

For the formal evaluation in experiment 1b, analysis of variance (ANOVA) will be used in a mixed-model form. This is due to the design of the test where one test participant will assess multiple parameter combinations, thus warranting the repeated measures model. In addition, the test is not perfectly balanced due to test organization practicalities. Thus, a between-subject factor is required and a mixed-model analysis of variance is used.

In addition, before any repeated measures model can be applied, the assumption of sphericity has to be tested. Mauchly's test of sphericity is used for this purpose. If it reveals any significant effects for intended factors and interactions of ANOVA, sphericity cannot be assumed in these cases. Nevertheless, correction can be applied to the degrees of freedom used in the further analysis in these cases, based on the value of sphericity ε. If $\varepsilon < 0.75$, the Greenhouse-Geisser correction is applied, and if $\varepsilon >= 0.75$, the Hyunh-Feldt correction is applied. After these corrections, the F-values of the ANOVA analysis are valid. [28]

3.5 Results of Experiment 1a—Evaluation of Spatial Perception

The main information provided by the participants in this experiment is their perceptual estimates of the distribution of sound into different azimuth directions for each parameter combination. Their answers are pooled in two circular histograms providing an estimate of the distribution of any (i.e., *minor* or *major*) sound and *major* sound for each parameter combination. This results in 96 different plots containing the histograms. However, as the histograms are quite similar between different program material, they are combined here for presentation purposes. Separated data is provided in the additional material online (see Sec. 8).

Fig. 6 shows the histograms of the combined data, and Fig. 7 visualizes the total number of marked loudspeakers (i.e., the blue lines in Fig. 6) for each parameter combination. The values in these figures are normalized so that 100% means that all participants answered that direction for all program material.

Fig. 7 reveals several central properties that agree with the informal experiments performed by the authors. First, the mirror distribution method improves the perceived distribution greatly when the window size is small, and mirror distribution does not seem to depend on the window size within the parameter values used. Second, the Halton sequence improves the perceived distribution, although less than was expected. Third, decorrelation makes the Halton sequence distribution more surrounding. Finally, the clear effect is that increasing the window size reduces the differences between different distribution methods and processing techniques.

In addition to these effects, a number of interesting detailed effects can be seen. Fig. 6 shows that the Halton sequence is directive in the unprocessed and decorrelated 256-sample window length cases. Furthermore, it is evident that the perceived distribution is even and does not have any clear *major* directions in the better parameter combinations (i.e., in all of the mirror distribution cases and most of the 1024- and 2048-sample window length cases). On the other hand, the cases on the top right of Fig. 6 that use a combination of uniform random distribution, short window size, and do not use mirror distribution, do not produce a desired proper spatial extent. This is evident due to the fact that they only achieve a 50% histogram value, which indicates that, on average, there were half of the possible directions missing for each listener. In these cases, the individual responses were quite random, containing both surrounding and narrow response patterns. These results agree with what was predicted to happen. Based on informal listening, a combination of random distribution with a short window size can often produce uneven distribution samples.

Another aspect visible in Fig. 6 is the bars on the right of each histogram, indicating the percentage of times that subjects had a perception of the sound to be inside the head or above or below the horizontal plane. Perception inside the head is not desired, and since there are a number of cases with shorter window lengths where such indications are common, those cases are deemed to be undesirable. Perception above or below, on the other hand, might be a positive effect if it is accompanied with an otherwise surrounding perception, as it makes the perception even more surrounding. Interestingly, mirror distribution generally produces more

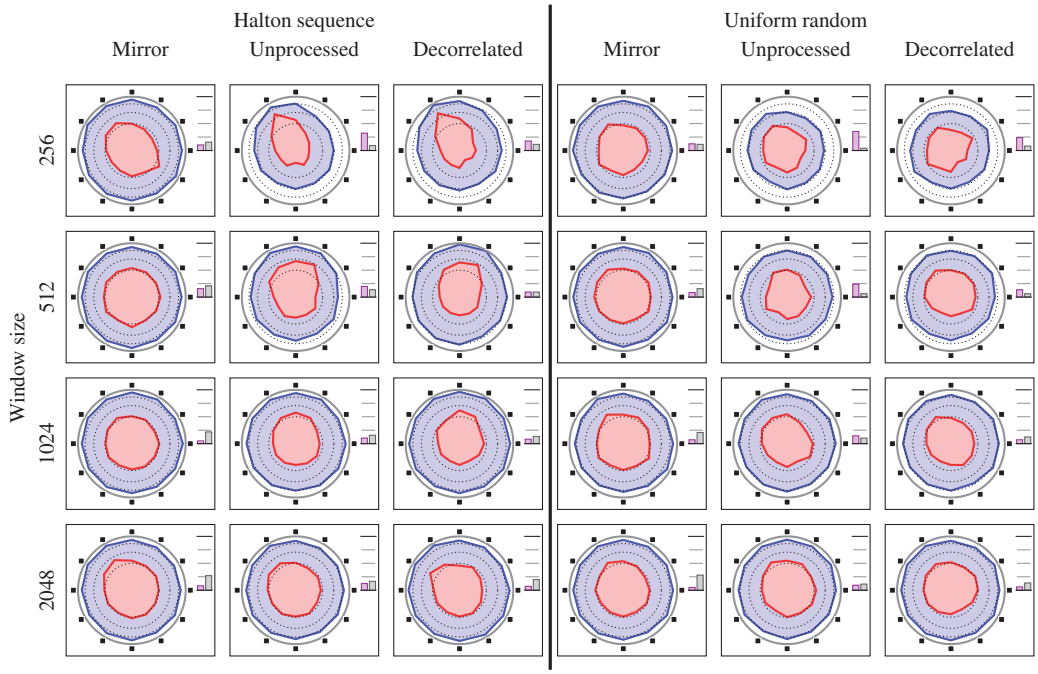

Fig. 6. Distribution histograms for the combined program material data. Blue represents any sound (*minor* or *major*) and red represents *major* sound. The thick gray line is the 100% marker and the dotted lines are 75%, 50%, and 25% markers. The histogram is on a square-root scale, making areas visually comparable. Small black boxes indicate the loudspeaker directions. The bars on the right side of each histogram represent the percentage of subjects answering that the sound comes from inside or near the head (left bar, violet) and from above or below (right bar, gray).

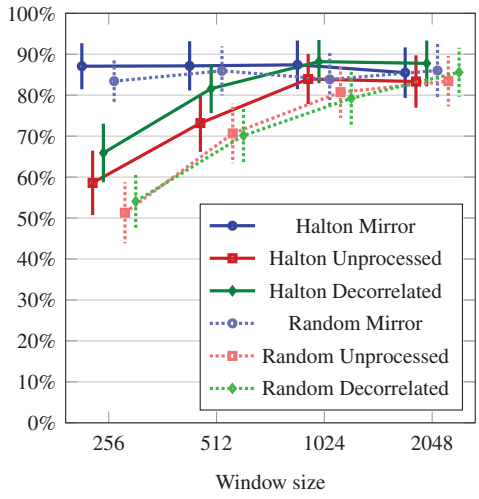

Fig. 7. Percentage of loudspeakers marked as the direction of the auditory event for each parameter combination and the 95% confidence intervals.

indications of perception above or below than the other parameter combinations.

The verbal and written comments collected during the experiment from the test subjects give additional insight into the perception of the reproductions. The discrimination between the *major* and *minor* directions were reported to be done so that when something exceptional was perceived from a direction, it was marked as a *major* direction. This agrees with the instructions shown in Sec. 3.1. The exceptional information could be either a higher sound pressure level, an audible coloration, localizable artifacts, or generally a significant point-like part of the sound without the adjacent directions perceived to be emitting sound. *Minor*, on the other hand, indicated that sound was perceived to be at the same level in the corresponding directions, or the sound was described as being diffuse. Typically, there was either only a few or no *major* directions, whereas the others were *minor*. Furthermore, some test subjects noted that in almost all the cases, they perceived sound to be emitted from all directions and that indication of no sound was rare. Interestingly, on some occasions, the cases with the cello signals were found to be divided so that specific frequencies, or changing pitch in the melody, were emitted from one direction per each note. This indicates that changing pitch may aid in finding new *major* directions, and that in musical sounds the divided frequency bands may be dominant in some directions, and therefore a prominent part of the auditory event forms there.

The importance of rotation in assessing the perceived sound is supported by the comments stating that on many occasions, the perceived sound scenario clearly changed from the initial impression after the participant rotated. Some participants reported that the sound tended to be perceived out of eyesight: initially, sound was perceived from the back, but when turning towards that direction, the perception shifts to another direction. This supports the idea of having a sound with no apparent main direction. On the other hand, such an effect also implies that the auditory event changes with the listener rotation, which was not a desired property.

Table 2. Significant effects in the mixed model ANOVA analysis.

Source	F	Sig.
program material	$F(3, 66) = 7.279$	0.000
window size	$F(1.605, 35.308) = 21.893$	0.000
distribution method	$F(1, 22) = 9.262$	0.006
processing technique	$F(2, 44) = 67.749$	0.000
program material * window size	$F(9, 198) = 3.576$	0.000
program material * processing technique	$F(6, 132) = 44.477$	0.000
window size * processing technique	$F(6, 132) = 2.207$	0.046
distribution method * processing technique	$F(2, 44) = 3.480$	0.040
window size * processing technique * group	$F(6, 132) = 3.842$	0.001
program material * distribution method * processing technique	$F(6, 132) = 3.177$	0.006
window size * distribution method * processing technique	$F(6, 132) = 3.812$	0.002

3.6 Results of Experiment 1b—Evaluation of Subjective Preference

The preference scores given by the participants were conservative, such that the extremes were not used often. This was anticipated, as the preference scale was defined with two quite extreme statements and does not create any problems. It should be emphasized here that the given preference scores do not evaluate the performance of the spatial extent synthesis, but they evaluate the overall preference of the samples including all the perceivable qualities.

Mixed-model ANOVA was performed to analyze the preference scores of experiment 1b. The within-subject factors were *program material*, *window size*, *distribution method*, and *processing technique*. The between-subject factor was *group*, which represents a participant starting with experiment 1a or 1b.

Mauchly's test of sphericity revealed two significant effects: (1) factor *window size*, $\chi^2(5) = 32.366$, $p < 0.05$, $\varepsilon < 0.75$; and (2) interaction *program material * window size * distribution method * processing technique*, $\chi^2(170) = 257.314$, $p < 0.05$, $\varepsilon < 0.75$. Correction for the degrees of freedom were applied in these cases in further analysis, as explained in Sec. 3.4. As $\varepsilon < 0.75$ in both of these cases, Greenhouse-Geisser correction was used.

The 11 different significant effects in ANOVA analysis are shown in Table 2. Further inspection of these effects was performed, and the most interesting effects are plotted in Figs. 8, 9, and 10. Plots of the other significant effects are provided in the additional material online (see Sec. 8). The plots shown here do not have any compensations applied for multiple comparisons, and thus these figures should be inspected for trends and not used to decide statistical significances.

Starting from the highest order significant interaction (found last in Table 2), *window size * distribution method * processing technique*, the effect in this case is minor. Mirror distribution is less affected by window size in random distribution, and random distribution is preferred overall. The interaction *program material * distribution method * processing technique* shows a small reduction in preference score in the unprocessed sea program material from the Halton sequence to random distribution. The only significant between-subject effect is in the interaction *window size * processing technique * group*. It shows that the group that started with experiment 1b preferred less the unprocessed signal with shorter window lengths. The only effect in the interaction *window size * processing technique* is that the unprocessed signal with a window size of 256 samples received a lower preference score.

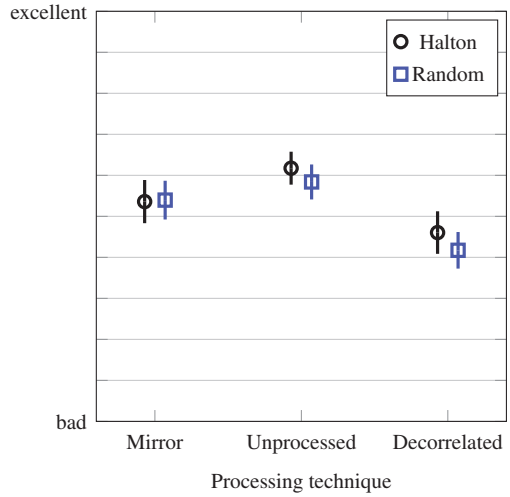

Fig. 8. Marginal means and 95% confidence intervals for the interaction between *distribution method* and *processing technique*.

The interaction *distribution method * processing technique* is shown in Fig. 8. In this case, it is interesting that the preference score does not degrade from the Halton sequence case to the random distribution case when mirror distribution is used, causing the interaction effect. This again supports the other results that mirror distribution makes the distribution less sensitive to other parameters.

In Fig. 9, the effect of the interaction *program material * processing technique* is shown. The main effect is in the speech program material. It is affected significantly more by the processing technique than the other program materials. The unprocessed signal received much better preference scores. This is due to the fact that the decorrelation artifacts are clearly audible in the speech program material with the two other processing techniques. In addition, this result creates doubt whether the results of different program materials are comparable. It is possible that the grading scale is constant only within one program material. This is

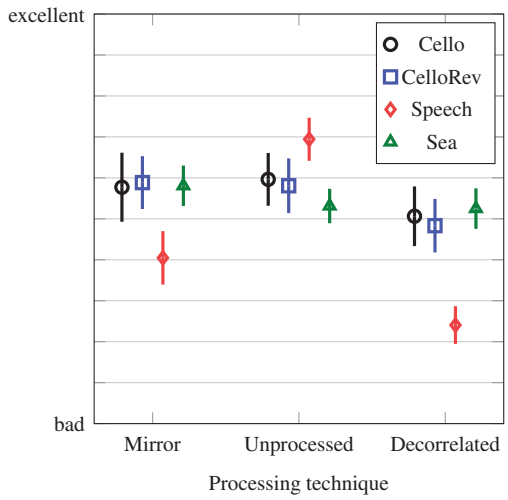

Fig. 9. Marginal means and 95% confidence intervals for the interaction between *program material* and *processing technique*.

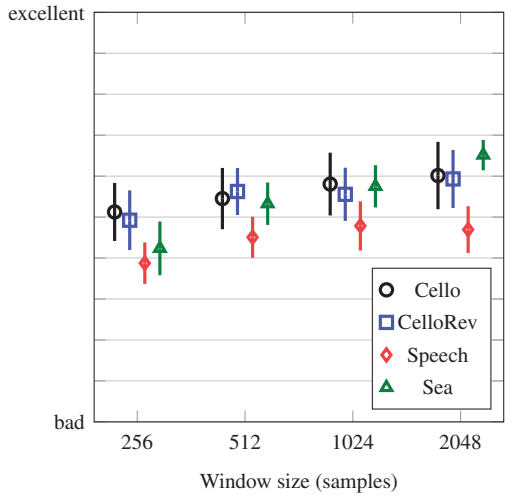

Fig. 10. Marginal means and 95% confidence intervals for the interaction between *program material* and *window size*.

reasonable, as the experimental organization separated each of the program materials into their own grading sessions. Thus, direct comparison of preference scores between program materials is not advised.

The final significant interaction *program material * window size* is shown in Fig. 10. There are two effects present in this figure. The first effect is that speech receives a worse preference score overall, although this should be taken with a grain of salt based on the results in the interaction *program material * processing technique*. The second effect is that the seashore program material improves more with the increase in window size than the other program materials.

The main effects of ANOVA did not reveal much more than has been already mentioned in the interactions, although these results do tell the most straightforward information for applications. Speech program material has a worse preference score than the other program materials, i.e., signal content affects quality. An increase in window size improves the preference score with the used parameter values. The Halton sequence is slightly better than the uniform random distribution. And finally, unprocessed is better than mirror distribution, which in turn is better than decorrelated processing.

Generally, giving subjective preference scores in experiment 1b was reported to be easier than indicating the directions in experiment 1a. As noted, in addition to giving the score, there were two attributes that could be marked as affecting the score positively or negatively. In the comments, naturalness was mentioned to be affected negatively when the stimulus did not sound as one would expect, and specifically, when the timbre was affected negatively. Spatial impression was more signal-dependent. There were indications that the seashore would optimally be all around the listener, whereas speech should be more prominently in one direction.

There were several comments regarding the speech test cases. The presented cases were noted to be belonging to two distinct groups, one described as having artifacts, a chorus-effect, unnaturalness, or overall annoyingness, the other group being the opposite. The cases belonging to the former group received lower preference scores. In terms of spatial impression, many noted that speech coming from only one direction would be more desirable than spatially surrounding. Overall, the comments reflected that the presented method is not easily suitable for processing speech and results in unnatural perception.

The cello in short reverberation was reported to be the most suitable for this kind of presentation and generally received the fewest descriptions of the sound being unnatural. Compared to the anechoic cello, it felt more believable to have a surrounding perception, and it was also mentioned that the reverberation seemed to be more “forgiving” for the reproduction. Reported unnatural effects on the anechoic cello test cases included a feeling of blocked ears and changing of perceived directions when rotating.

In the case of the seashore, some test subjects noted hearing a sort of Doppler effect when rotating in the cases where they perceived surrounding sound. This is likely caused by the division of different frequencies of noise-like sound into different directions. In addition, there was an interesting comment that some cases caused a feeling of being inside a large swimming pool, indicating immersion in the scenario.

4 FORMAL EVALUATION—EXPERIMENT 2

The aim of the second part of the formal evaluation was to find out how the presented method can produce different spatial extents. The experimental organization and the results of the second experiment are presented in the following sections.

4.1 Experiment Design—Evaluation of Different Spatial Extents

Only one parameter combination was used in this experiment, and it was selected based on the combined results

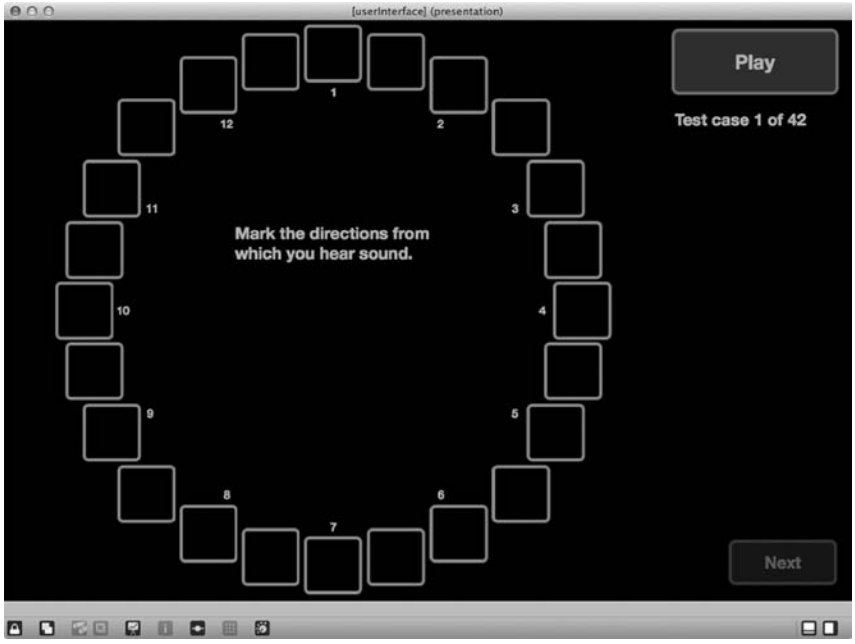

Fig. 11. User interface for experiment 2.

of experiment 1 and informal experiments. This combination used the Halton sequence for distribution, 1024-sample window size, and the unprocessed synthesis. This combination proved to be very suitable for all used signal types with no adverse artifacts.

Cello music in short reverberation and the recording at a seashore were used as the program material in this experiment. Furthermore, discrete fully uncorrelated pink noise was used as the third program material. This was done to give a reference that would theoretically have the widest possible perceived spatial extent. Furthermore, this enables comparison to a previous study by Santala and Pulkki [13].

Seven different extents were produced in this experiment. These were symmetrical sets with 1, 3, 5, 7, 9, 11, or 12 loudspeakers corresponding to 0°, ±30°, ±60°, ±90°, ±120°, ±150°, or ±180° spatial extent. Two repetitions were performed for each extent.

The task of the participants was to mark, using the provided GUI (shown in Fig. 11), the directions from which they perceived any sound. In this case the GUI also provided directions in the middle between the actual loudspeaker positions to offer more granularity for answers. Furthermore, participants could write comments on a paper and give verbal comments to the organizer.

4.2 Experiment Organization

Experiment 2 was organized in a similar manner as the first experiment with only a few differences. Fifteen of the 24 participants in the first experiment participated in experiment 2 as well. The experiment consisted of a single session with one break. As the participants assessed each parameter combination twice, there were 42 cases to be assessed in total. The direction of arrival of the produced extent was randomized separately for each test case and participant. Otherwise, the organization was the same as in experiment 1. On average, each participant took 40 to 50 minutes to complete experiment 2.

4.3 Results of Experiment 2—Evaluation of Different Spatial Extents

The results of experiment 2 reveal how the presented method can produce different perceptual spatial extents. Fig. 12 shows the circular histograms of the marked directions in relation to the used loudspeakers. There are several effects present in these results.

The natural effect is that the perceived extent becomes larger as the area used for reproduction becomes larger. However, with the presented method (i.e., “Cello” and “Sea” rows in the figure), the perceived extent is not as wide as the extent of the corresponding loudspeaker setup until a fully surrounding case is presented. With the reference pink noise (“Pink” row in the figure), the perceived extent is almost as wide as the loudspeaker setup used. Finally, there seems to be some tendency to the left side with the presented method with extents from ±60° to ±120°. It should be noted that this cannot be an effect caused by the experimental setup itself, since the presentation direction was always randomized.

The participants expressed a general feeling that experiment 2 was easier than the previous ones, indicating that when only some of the loudspeakers were emitting sound, the auditory event was easier to analyze. This is logical since the nature of the method was to produce ambient sound in the case where all loudspeakers were used, and analyzing such scenarios can be considered to be harder than scenarios where only a few loudspeakers are emitting sound. In addition, more complex distributions are harder to perceive accurately [13].

Overall, the results show that the presented method can be used to synthesize different perceptual spatial extents,

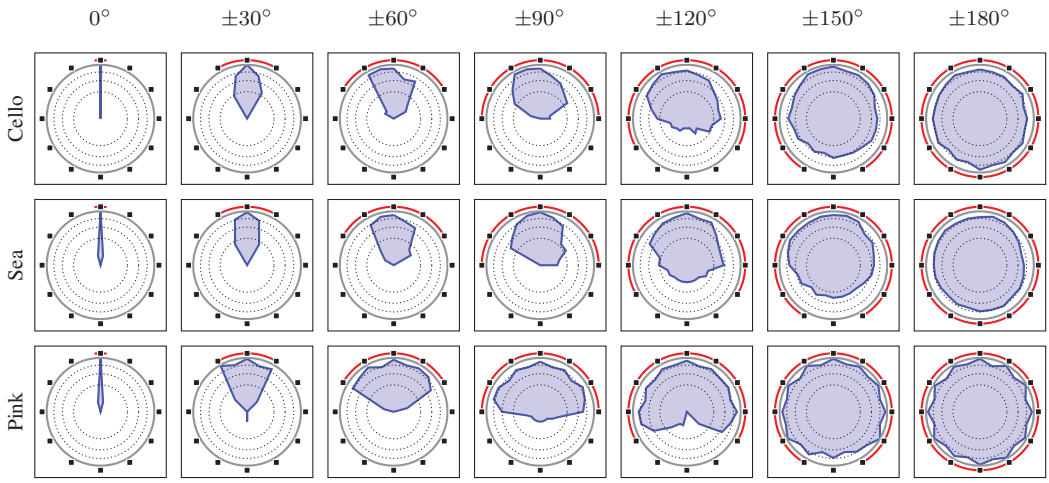

Fig. 12. Distribution histograms for different widths. The red line represents the area which contains the used loudspeakers, while the small black boxes indicate the directions of the loudspeakers. The thick gray line is the 100% marker, and the dotted lines are 75%, 50%, and 25% markers. The histogram is on a square-root scale, making areas visually comparable. Note that it was possible to also answer between the shown loudspeakers.

although the mean result is not as wide as the used loudspeaker setup.

5 DISCUSSION

In this section the implications and questions raised by the results are discussed.

5.1 Implications of Results

Based on the results, the revised method does indeed fulfill its task—it creates perceptual spatial extent for a monophonic source signal. Furthermore, it can synthesize different spatial extents successfully. However, there are several significant implications that affect the user if they want to apply the revised method.

First and foremost, the results of the formal experiments suggest the same as was theorized based on the informal experiments—signal content affects the output quality of the revised method. This means that the method cannot be used blindly to process all signals, as some signal types become perceptually undesirable (e.g., inside the head) if processed with the method. Nevertheless, if this property is taken into account, this method can be applied to most signals blindly and to certain signals with a proper selection of parameters. For signals with a mainly impulsive content, the use of other algorithms is advised.

The main results reveal what parameter combinations should be preferred (although most were plausible) when using the revised method. If a short and simple answer is desired, then the combination used in experiment 2 is the prime candidate. This means using a window size of 1024 samples and distributing the frequency bands using the Halton sequence. This parameter combination produces most often an output signal with a perceptually even and surrounding spatial extent with no adverse artifacts. Furthermore, even better results can often be obtained by changing the parameters depending on the signal content.

One especially interesting topic is the performance of the mirror distribution, as it is a novel method. Based on the formal results, it certainly seems to solve many problems caused by other misaligned parameters. Additionally, the perceived spatial extent seems to be most surrounding when using this technique. However, the subjective preference scores show that it is not always preferred. This is due to the decorrelation artifacts being audible with certain signal content, although to a much lesser extent than with single decorrelation case. In experiment 1b, such dispreferred signal content was speech, as can be seen in Fig. 9. Thus, the mirror distribution method is very promising, and it is a very good choice for spatialization tasks.

As for the synthesis of different spatial extents (i.e., the results of experiment 2), the revised method seems to produce deterministically a perceptual spatial extent when a specific extent is desired. However, the mapping is not one-to-one and linear. Instead, the perceived extent is narrower than intended until the intended extent is almost surrounding the listener. Nevertheless, although not done in this paper, it should be possible to create an appropriate mapping function from intended spatial extent to perceived spatial extent with the help of the presented results and further experiments if that is desired for practical applications. Interestingly, the pink noise reference produced a wider perceived spatial extent than the signals processed with the revised method.

5.2 Perceptual Aspects of Results

The presented research evoked observations and questions that are interesting from a psychoacoustical and perceptual perspective. The main observation is that distributing frequency bands to different spatial locations yields a spatially extended perception. Furthermore, this perception

often does not have apparent directional focus. Interestingly, this suggests that no spatial summation occurs, but perceptual fusion of timbre is still obtained.

The revised method distributed narrow frequency bands in different directions so that adjacent frequency bands would not be close to each other. The suitability of this approach for creating a perceptually spatially spread auditory event is supported by the studies of Hirvonen and Pulkki [16]. As discussed in Sec. 1.2, they found that when adjacent frequency bands were not presented from neighboring loudspeakers, the perceived width was wider than when having the frequency bands spatially next to each other. The results obtained in the present experiments are in line with this finding.

Furthermore, because a significant part of the test cases were perceived to be almost everywhere around the listener, it can be said that in such cases the frequency bands were not spatially fused. However, as shown in Fig. 12, the different spatial extents produced with the revised method in experiment 2 were found to be perceptually narrower than the width of the sound source constellation. When the loudspeaker span was $\pm 120^\circ$ or less, the perceived extent was close to being only half of that of the loudspeaker setup. This was the case in [15] as well, and in that study, the width of the frequency band in each loudspeaker was one ERB, indicating that there are similarities in the perception when spreading very narrow frequency bands or ERB bands. However, the informal listening in the present study showed that when using ERB bands in a completely surrounding setup, the resulting auditory event is not evenly spread, whereas with narrow frequency bands, such a perception was formed. The reasons for these observations need further research.

In experiment 2, the results of the four test cases where pink noise was presented from 0° to $\pm 90^\circ$ can be compared to the results of the experiment by Santala and Pulkki [13], mentioned in Sec. 1.2. Compared to that experiment, the main difference in the test procedure of the present experiment is that the loudspeakers were placed every 30° instead of 15°. The results of these experiments are very close to each other, although, in the present experiment, the perceived extent was closer to the intended extent. Nevertheless, the alignment of results suggest that the results of the present experiment are plausible and provide information from the psychoacoustic perspective as well.

Another interesting observation based on the results of experiment 2 is that in the case of the widest distribution width, $\pm 180^\circ$, the revised method produces a more coherent and even perception than the incoherent pink noise. In addition, some test subjects specifically noted that, in the case of the pink noise with distribution widths of $\pm 150^\circ$ and $\pm 180^\circ$, they perceived sound to come from the loudspeaker directions but not from between them. The effect is comparable to the headphone studies on IACC discussed in Sec. 1.2 [10], where it was found that high IACC resulted in a wide auditory event inside the head, while low IACC resulted in separated auditory events inside the head. In the present study the pink noise signals were incoherent and could therefore facilitate the perception of separate auditory events, similarly as in the case of IACC studies. Furthermore, with incoherent signals, the temporal envelopes may have distinct peaks at different time instants, thus making the different signals stand out from the overall scenario. However, in [13], the perception of a loudspeaker ensemble with a span of $\pm 90^\circ$ and 30° spacing was not significantly different from that of an ensemble with 15° spacing. The presence of visual cues in 30° intervals in the present experiment may have influenced the perception. However, no such effects occurred with the test cases of the revised method, and this effect needs further research.

6 SUMMARY

This paper revised an existing method for synthesizing spatial extent for a monophonic input signal. The method is performed in the time-frequency domain and distributes frequency bands randomly or deterministically into different spatial locations. The original method could synthesize spatial extent but not always reliably or with good timbral quality. The method was now reformulated and inspected properly to understand how it is perceived and why it is perceived as it is. This revised method achieves reliable synthesis of spatial extent with good timbral quality. Furthermore, the presented formulation is independent of the reproduction technique.

A series of informal and formal experiments were organized to study the properties of the presented method and verify its suitability. It was indeed found that the method is very suitable for synthesizing spatial extent for monophonic input signals. However, the resulting quality is signal-dependent. This mainly means that sparse and impulsive signals should be processed with other methods than the presented one. On the other hand, with a suitable continuous signal and a proper parameter combination, the resulting sound scene is perceived as very surrounding and even.

As for the proper parameter combination, there is no such combination that would always be the best due to signal dependency. Nevertheless, the suggested choice in most situations is to use an STFT window size of 1024 samples and then distribute the frequency bands from lowest to highest using the Halton sequence.

From a psychoacoustical perspective, the presented study evoked new perspectives and questions on the perception of spatial extent. They all concentrate around the central result of the method – spatially distributed narrow frequency bands of a monophonic source can create a single unified spatially extended perception.

7 ACKNOWLEDGMENTS

This work has been supported by the Academy of Finland. The research leading to these results has received funding from the European Research Council under the European Community's Seventh Framework Programme (FP7/2007-2013) / ERC grant agreement n^o [240453].

8 ADDITIONAL MATERIAL

Additional results of the experiments and binaural renderings are provided on the companion webpage http://www.acoustics.hut.fi/go/jaes-extent/.

9 REFERENCES

[1] P. D. Coleman, "Failure to Localize the Source Distance of an Unfamiliar Sound," *J. Acous. Soc. Am.*, vol. 34, no. 3, pp. 345–346 (1962).

[2] V. Pulkki, M.-V. Laitinen, and C. Erkut, "Efficient Spatial Sound Synthesis for Virtual Worlds," *Audio Engineering Society 35th International Conference*: Audio for Games (2009 Feb.), paper 21.

[3] M.-V. Laitinen, T. Pihlajamäki, C. Erkut, and V. Pulkki, "Parametric Time-Frequency Representation of Spatial Sound in Virtual Worlds," *ACM Transactions on Applied Perception*, vol. 9, no. 2 (2012).

[4] C. Verron, M. Aramaki, R. Kronland-Martinet, and G. Pallone, "A 3-D Immersive Synthesizer for Environmental Sounds," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 6, pp. 1550–1561 (2010).

[5] G. Potard and I. Burnett, "Decorrelation Techniques for the Rendering of Apparent Sound Source Width in 3D Audio Displays," *Proceedings of the 7th International Conference on Digital Audio Effects (DAFx-04)*, pp. 280–284, Naples, Italy (2004).

[6] F. Zotter, M. Frank, M. Kronlachner, and J.-W. Choi, "Efficient Phantom Source Widening and Diffuseness in Ambisonics," *Proceedings of the EAA Joint Symposium on Auralization and Ambisonics*, Berlin, Germany (2014).

[7] P. Pestana and J. Reiss, "A Cross-Adaptive Dynamic Spectral Panning Technique," *Proceedings of the 17th International Conference on Digital Audio Effects (DAFx-14), Erlangen, Germany* (2014).

[8] R. Y. Litovsky, H. S. Colburn, W. A. Yost, and S. J. Guzman, "The Precedence Effect," *J. Acous. Soc. Am.*, vol. 106, no. 4, pp. 1633–1654 (1999).

[9] J. Blauert, *Spatial Hearing* (The MIT Press, Cambridge, MA, USA, revised edition, 1997) pp. 1–494.

[10] J. Blauert and W. Lindemann, "Spatial Mapping of Intracranial Auditory Events for Various Degrees of Interaural Coherence," *J. Acous. Soc. Am.*, vol. 79, no. 3, pp. 806–813 (1986).

[11] R. Mason, T. Brookes, and F. Rumsey, "Frequency Dependency of the Relationship between Perceived Auditory Source Width and the Interaural Cross-Correlation Coefficient for Time-Invariant Stimuli," *J. Acous. Soc. Am.*, vol. 117, no. 3, pp. 1337–1350 (2005).

[12] T. Hirvonen and V. Pulkki, "Perceived Spatial Distribution and Width of Horizontal Ensemble of Independent Noise Signals as Function of Waveform and Sample Length," presented at the 124th Convention of the Audio Engineering Society (2008 May), convention paper 7408.

[13] O. Santala and V. Pulkki, "Directional Perception of Distributed Sound Sources," *J. Acous. Soc. Am.*, vol. 129, no. 3, pp. 1522–1530 (2011).

[14] D. R. Perrott and T. N. Buell, "Judgments of Sound Volume: Effects of Signal Duration, Level, and Interaural Characteristics on the Perceived Extensity of Broadband Noise," *J. Acous. Soc. Am.*, vol. 72, no. 5, pp. 1413–1417 (1982).

[15] T. Hirvonen and V. Pulkki, "Center and Spatial Extent of Auditory Events as Caused by Multiple Sound Sources in Frequency-Dependent Directions," *Acta Acustica united with Acustica*, vol. 92, pp. 320–330 (2006).

[16] T. Hirvonen and V. Pulkki, "Perception and Analysis of Selected Auditory Events with Frequency-Dependent Directions," *J. Audio Eng. Soc.*, vol. 54, pp. 803–814 (2006 Sep.).

[17] K. Hiyama, S. Komiyama, and K. Hamasaki, "The Minimum Number of Loudspeakers and its Arrangement for Reproducing the Spatial Impression of Diffuse Sound Field," presented at the *113th Convention of the Audio Engineering Society* (2002), convention paper 5696.

[18] F. Rumsey, "Spatial Quality Evaluation for Reproduced Sound: Terminology, Meaning, and a Scene-Based Paradigm," *J. Audio Eng. Soc.*, vol. 50, pp. 651–666 (2002 Sep.).

[19] J. Berg, "The Contrasting and Conflicting Definitions of Envelopment," presented at the 126th Convention of the Audio Engineering Society (2009 May), convention paper 7808.

[20] L. Beranek, Concert and Opera Halls: How They Sound, published for the Acoustical Society of America through the American Institute of Physics (1996).

[21] D. Griesinger, "The Psychoacoustics of Apparent Source Width, Spaciousness and Envelopment in Performance Spaces," *Acta Acustica united with Acustica*, vol. 83, no. 4, pp. 721–731 (1997).

[22] J. Ahrens and S. Spors, "Two Physical Models for Spatially Extended Virtual Sound Sources," presented at the *131st Convention of the Audio Engineering Society* (2011 Oct.), convention paper 8483.

[23] B. R. Glasberg and B. C. J. Moore, "Derivation of Auditory Filter Shapes from Notched-Noise Data," *Hearing Research*, vol. 47, pp. 103–138 (1990).

[24] M.-V. Laitinen, F. Kuech, S. Disch, and V. Pulkki, "Reproducing Applause-Type Signals with Directional Audio Coding," *J. Audio Eng. Soc.*, vol. 59, pp. 1–15 (2011 Jan./Feb.).

[25] J. H. Halton and G. B. Smith, "Radical-Inverse Quasi-Random Point Sequence," *Communications of the ACM*, vol. 7, no. 12, pp. 701–702 (1964).

[26] M.-V. Laitinen, S. Disch, and V. Pulkki, "Sensitivity of Human Hearing to Changes in Phase Spectrum," *J. Audio Eng. Soc.*, vol. 61, pp. 860–877 (2013 Nov.).

[27] M. Bouéri and C. Kyriakakis, "Audio Signal Decorrelation Based on a Critical Band Approach," presented at the 117th Convention of the Audio Engineering Society (2004 Oct.), convention paper 6291.

[28] A. Field, Discovering Statistics Using SPSS, ISM Introducing Statistical Methods (SAGE Publication Ltd., London UK, 2005), 2

THE AUTHORS


Tapani Pihlajamäki


Olli Santala


Ville Pulkki

Tapani Pihlajamäki began his work on acoustics and audio signal processing in 2003 when he was accepted into Helsinki University of Technology. He completed his Master's degree in 2009 majoring in acoustics and audio signal processing. He continued his studies towards a Doctoral degree in the university's (now called Aalto University) department of signal processing and acoustics under the guidance of Ville Pulkki. His thesis concentrates on developing methods for spatial audio synthesis and reproduction in virtual realities. Currently, he is finishing his work on the degree and extensively ponders what might happen afterwards.

•

Olli Santala received his M.Sc. (Tech) degree in 2009 from former Helsinki University of Technology, now called Aalto University, majoring in acoustics and cognitive technology. Currently he is pursuing his D.Sc. at Aalto University. The focus of his research is on spatial sound perception, approached by conducting psychoacoustic listening experiments, modeling the auditory pathway, as well as collaborating in brain studies on hearing using magnetoencephalography. Music-related free-time activities include singing in a vocal ensemble and playing the guitar.

•

Ville Pulkki has worked in the field of spatial audio since 1995. In his Ph.D. thesis (2001) he developed a method to position virtual sources for 3-D loudspeaker set-ups and researched the method using psychoacoustic listening tests and binaural computational models of human hearing. Later, he also worked on reproduction of recorded spatial sound scenarios, on measurement of head-related acoustics, and on measurement of room acoustics with laser-induced pressure pulses. Currently he holds a tenure-track assistant professor position at Aalto University and runs a research group with 18 researchers. He also has a background in music, having been taught various instruments at Sibelius-Academy, in singing, and in audio engineering. He has also composed and arranged music for different ensembles. He enjoys being with his family, renovating his summerhouse, and dancing hip hop.