HIGH ANGULAR RESOLUTION PLANEWAVE EXPANSION

Svein Berge

Berges Allmenndigitale
Darres gate 20, N-0175 Oslo, Norway
sveinb@pvv.org

Natasha Barrett

Blåsbortveien 10, N-0873 Oslo, Norway
nlb@natashabarrett.org

ABSTRACT

First-order ambisonics suffers from low angular resolution and a small sweet spot, and decoding to many loudspeakers does not help. Parametric decoding methods solve this problem, at the risk of introducing audible artifacts. A method for high angular resolution planewave expansion (HARPEX) is proposed, which combines the spatial sharpness of parametric methods with the physical correctness of linear decoding without introducing audible artifacts. In a formal listening test, a decoder using this method to decode first-order B-format signals scores much higher than max r_E decoding of the same signals, and similarly to max r_E decoding of third-order versions of the same signals.

1. INTRODUCTION

When reproducing a B-format signal [1], the angular resolution and size of the sweet spot depends on the number of loudspeakers and on the number of input channels. Increasing the number of loudspeakers without also increasing the B-format order does not improve the angular resolution with linear decoding methods [2].

Previous work on sharpening the spatial image [3, 4, 5] has split the signal into frequency bands and used the short-time correlation between the W channel and each of the directional channels to calculate an estimate of the direction of arrival and "diffuseness" of the sound in each frequency band, which in turn has been used to steer part of the signal to the loudspeakers closest to that direction. The difference between these different methods is mainly in their processing of the diffuse part of the signal.

Linear decoding methods are based on the decomposition of the sound field in planewaves or spherical waves emanating from the loudspeakers. A minimum of four planewaves is necessary to reconstruct the first-order sound field. However, as we shall see, if the direction of the planewaves is allowed to adapt to the signal, then two planewaves are sufficient to reconstruct the sound field.

The method proposed in this paper is based on the idea of decomposing each frequency component of the sound field in two planewaves and then reconstructing those planewaves with the available loudspeakers. The result is a physically correct reconstruction of the sound field at the sweet spot and a high-resolution planewave expansion outside it. At frequencies where only one or two planewaves contribute significantly to the recorded signal, this is a physically correct expansion. At frequencies with two planewaves, determining the correct directions of arrival pushes the capability of human hearing. With more than two planewaves at the same frequency, one may expect directional errors to go unnoticed.

2. PARAMETRIC DECOMPOSITION

The method operates in the frequency domain. The first step is to apply overlapping window functions, zero padding and FFT. Eight numbers then represent each time/frequency bin of a 3D first order signal: The real and imaginary part of each channel. This should be possible to decompose into two planewaves, each represented by four independent numbers: The real and imaginary part of the amplitude and a three-element unit vector pointing in the direction of arrival.

In the rest of this section, only one frequency component will be considered, assuming that the same method is applied to all frequency components. If $\mathbf{X}$ is the complex-valued signal, then the decomposition can be written as

$$\mathbf{X} = \begin{bmatrix} w_r + iw_i \\ x_r + ix_i \\ y_r + iy_i \\ z_r + iz_i \end{bmatrix} = \underbrace{\begin{bmatrix} 2^{-\frac{1}{2}} & 2^{-\frac{1}{2}} \\ x_1 & x_2 \\ y_1 & y_2 \\ z_1 & z_2 \end{bmatrix}}_{\mathbf{V}} \underbrace{\begin{bmatrix} a_1 \\ a_2 \end{bmatrix}}_{\mathbf{A}}, \quad (1)$$

where the bottom three rows of $\mathbf{V}$ contain real-valued unit vectors pointing in the directions of arrival and $\mathbf{A}$ contains the complex amplitudes of those waves. The decomposition can be calculated in several ways. One way is to first find the phases of a_1 and a_2 with the following formula, some basic algebra to find their magnitude and finally a matrix inversion to find $\mathbf{V}$.

$$\mathbf{A} = \begin{bmatrix} m_1 & 0 \\ 0 & m_2 \end{bmatrix} \begin{bmatrix} c_1 & s_1 \\ c_2 & s_2 \end{bmatrix} \begin{bmatrix} 1 \\ i \end{bmatrix} \quad (2)$$

$$c_{1,2} = \sqrt{\frac{2r^2 - pq + p^2 \pm 2r\sqrt{r^2 - pq}}{(q-p)^2 + 4r^2}} \quad (3)$$

$$s_{1,2} = \frac{(q-p)c_{1,2} + p/c_{1,2}}{2r} \quad (4)$$

$$r = -2w_r w_i + x_r x_i + y_r y_i + z_r z_i \quad (5)$$

$$p = -2w_r^2 + x_r^2 + y_r^2 + z_r^2 \quad (6)$$

$$q = -2w_i^2 + x_i^2 + y_i^2 + z_i^2 \quad (7)$$

There are cases ($r^2 - pq < 0$) where the decomposition does not exist. In an isotropical noise field this concerns 1/4 of all samples. In real sound recordings, however, this percentage is lower, and low-energy frequency bands are over-represented. The energy contained in these frequency bands usually sums up to around 2 to 3 percent of the total energy. These cases must be

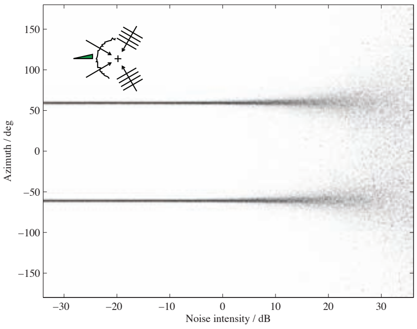

Figure 1: Direction estimates in the presence of noise

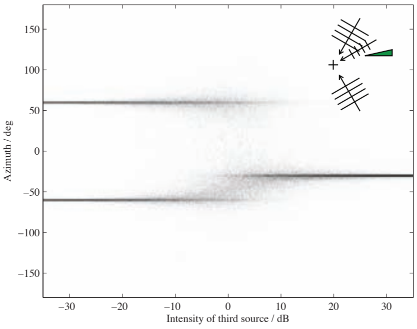

Figure 2: Directional noise

handled with an alternative method. Since they concern a small fraction of the signal, the choice of alternative method makes no perceptible difference, and these methods will not be treated further in this paper.

3. NUMERICAL EXPERIMENTS

The method is model based and non-linear, so it is not trivial to predict its behavior when the model assumptions are not valid. We will therefore show some numerical experiments that study its behavior under the influence of different types of noise, and close to its singularities. In each experiment, the results are visualized in a scatter plot where a single parameter is varied along the horizontal axis. The azimuth of the direction estimates is plotted along the vertical axis. For each horizontal pixel, 300 experiments are performed and two points are plotted for each experiment, unless the method fails to decompose the signal. The opacity of each point is proportional to the amplitude of the corresponding planewave.

3.1. Non-directional noise

In a first experiment, a signal is synthesized from two point sources of white noise. In addition, white noise is added to each of the four channels.

The dB scale has been shifted up by 36 dB, which corresponds to the signal-to-noise ratio of a sine tone and white noise of equal power, after filtering with a filter 2048 samples long, a typical window length used with HARPEX. This measure of signal-to-noise is comparable to SNR numbers.

The direction estimates are generally correct at noise levels below 0 dB and degrade gradually at noise levels above 10 dB.

3.2. Directional noise

In a second experiment, a signal is synthesized consisting of three point sources of white noise. The power of two of the sources remains constant while the power of the third, interfering source is varied over a range from –35 dB to +35 dB relative to each of the other sources.

The direction estimates are practically immune to interference below –20 dB. In a transition region where all three sources are in the same 10 dB power range, direction estimates fall in a broad region in the plane containing the three sources. When the interfering source is more than 20 dB stronger than the other sources, one direction estimate corresponds to the direction of the interfering source alone. A second direction estimate is also produced in this region, widely distributed over the plane. The amplitude corresponding to this second estimate is negligible, which is why it does not show in the figure.

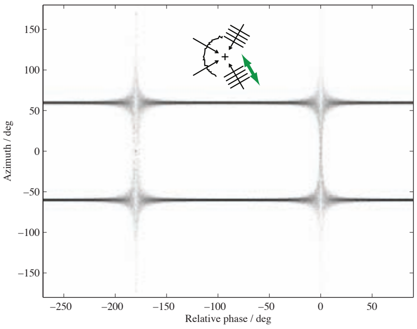

Figure 3: Phase degeneracy

3.3. Phase degeneracy

In a third experiment, a signal is synthesized consisting of two coherent, sinusoidal point sources. The relative phase of the sources is varied over a range from –180° to 180°. White noise is added to each channel, equal to 0 dB on the scale defined in the first experiment.

The direction estimates are generally correct except when the phase of the two waves differ by less than 10° or more than 170°, where direction estimates diverge. Other runs of the same experiment show that the regions of divergence narrows as the noise level decreases, while the amount of divergence remains constant.

3.4. Directional degeneracy

In a fourth experiment, a signal is synthesized from two point sources of equal power, one straight ahead and the other being moved in a circle around the horizontal plane. Noise is added at a level equal to 0 dB on the scale defined in the first experiment.

Direction estimates are generally correct for angles greater than 30° between the two planewaves. At smaller angles, direction estimates widen until the two sources are very close (< 5°). At this separation, the sources are fused into one direction estimate with correspondingly high amplitude.

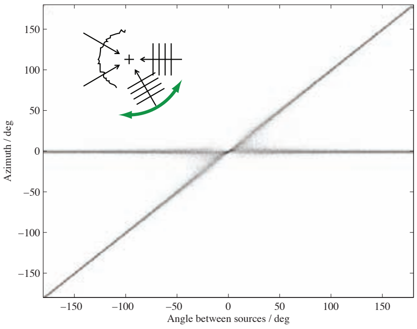

Figure 4: Directional degeneracy

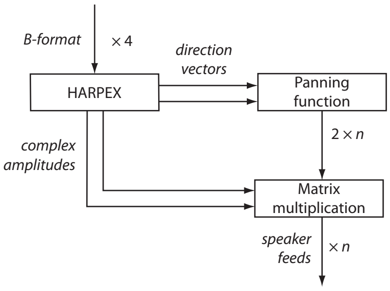

Figure 5: Simple decoder

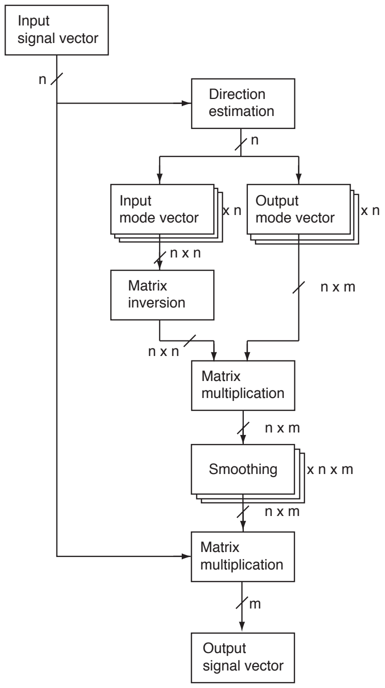

Figure 6: Complete decoder

4. DECODER IMPLEMENTATION

The most straightforward implementation of a decoder using HARPEX would be to send each of the two direction estimates into a panning function, which would return a weight for each of the output speakers. Each weight would then be multiplied with the complex amplitude of the corresponding planewave to generate speaker feeds. This approach is illustrated in Figure 5, not showing windowing, FFT and IFFT, which would also be necessary.

There are several problems with this implementation. Firstly, the HARPEX method does not always return a solution and must be accompanied with a fallback method to use in such cases. Of the known useful fallback methods, none provide more than a single direction estimate.

Secondly, the direction vectors may change rapidly from one frame to the next, causing time domain artifacts related to the frame period. One solution is to smooth the direction vectors, but a better solution is to smooth the resulting panning weights.

Thirdly, the direction vectors may differ significantly from one frequency bin to the next within a frame, causing dispersion and undesirably soft transients. Smoothing the direction vectors across the frequency axis can solve this problem, but again it is better to smooth the panning weights instead.

One effect of smoothing is to introduce leakage between sources that have been separated. For diffuse sources, this leakage is desirable, but for point sources, the amount of smoothing represents a trade-off between the sharpness of localization and the audibility of artifacts.

Whether the smoothing is done before or after the panning function, the decomposition into two plane waves is no longer valid, since the direction of the waves is altered. To regain a valid decomposition, another two planewaves can be added and the signal must be decomposed into this new basis. In cases where HARPEX returns no solution, three new planewaves must be added, so that the second decomposition always returns four waves.

To ensure good conditioning of the decoding matrix, the additional planewaves should be placed as far away as possible from the original planewaves and each other.

4.1. Panning functions

Since decomposition and resynthesis is split into two separate operations, any panning function can be used. The most obvious choice for horizontal loudspeaker layouts would be a pairwise panning, using a panlaw in the 3–6 dB range. This can easily support irregular layouts, and can be extended to with-height layouts using vector-based amplitude panning [6].

Other functions worth mentioning are ambisonics-equivalent panning functions [7] and wavefield synthesis [8]. Another interesting option is to use spherical harmonics as panning functions. The output of the decoder will in this case not be loudspeaker feeds, but rather an up-mix of first-order B-format to higher-order B-format. This panning function has the desirable property of reconstructing the sound field in the sweet spot, if combined with a suitable higher-order ambisonic decoder.

Decoding for binaural playback using head-related transfer functions presents additional challenges stemming from the fact that HRTFs contain phase terms that encode the interaural time delay. These phase terms can lead to audible artifacts unless further processing is undertaken. This will be the subject of a future publication.

5. LISTENING TEST

Since the proposed method aims to improve the perceived sound quality outside the region where the physical sound field can be reproduced, the only way to assess if it works according to expectations is to perform listening tests. We chose to follow as closely as possible the experimental setup of experiment no. 1 in [2], since this test provided clear and quantitative results. That experiment was in turn designed according to the MUSHRA recommendations [9].

5.1. Experimental setup

The test was carried out in an acoustically treated studio at NOTAM. Twelve Genelec 1030A loudspeakers were placed on a circle as shown in Figure 7. The speakers formed a regular octagon plus a standard ITU 5.0 layout, with the center speaker belonging to both sets.

Six sound scenes were used in the test, shown in Figure 8. A seventh scene was also created and only used for training (not shown). Two of the scenes (*enfant* and *cuisine*) were identical to scenes used in [2]. Since a possible weakness of any nonlinear method is the reproduction of scenes with multiple overlapping sounds, all scenes were chosen such that there were always at least three, usually more, overlapping sound sources. Each scene lasted between 10 and 17 seconds.

The reference signals consisted of one sound source per loudspeaker, routed to a subset of the twelve available loudspeakers. Across the six scenes, roughly half of the sources were placed on loudspeakers belonging to the octagon and half on loudspeakers belonging to the 5.0 layout.

Six systems were tested and compared to the reference:

1-8: 1st order, decoded with max r_E to octagon

3-8: 3rd order, decoded with max r_E to octagon

H-3-8: 1st order, up-mixed to 3rd order using HARPEX, 3rd order decoded with max r_E to octagon

H-8: 1st order, decoded with HARPEX and 3 dB pairwise panning to octagon

H-5: 1st order, decoded with HARPEX and 3 dB pairwise panning to ITU 5.0

REF: Hidden reference

In the H-3-8 system, the HARPEX panning function was equal to the spherical harmonics up to third order. The output from the HARPEX decoder was sent to the same decoder as the 3-8 system [10]. Participants were asked to rate each of the six signals in each of the six scenes on a scale from 0 to 100, associated with the following guidelines. The adjectives were given in English and the explanation in Norwegian.

80-100: "Excellent," no degradation

60-80: "Good," little change in position

40-60: "Fair," deviation from original position, sources widening

20-40: "Poor," substantial deviation from original position, sources widening, difficult to localize sound sources

0-20: "Bad," sources are completely out of their original position, very hard to localize.

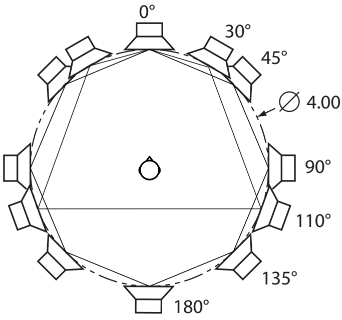

Figure 7: Loudspeaker layout

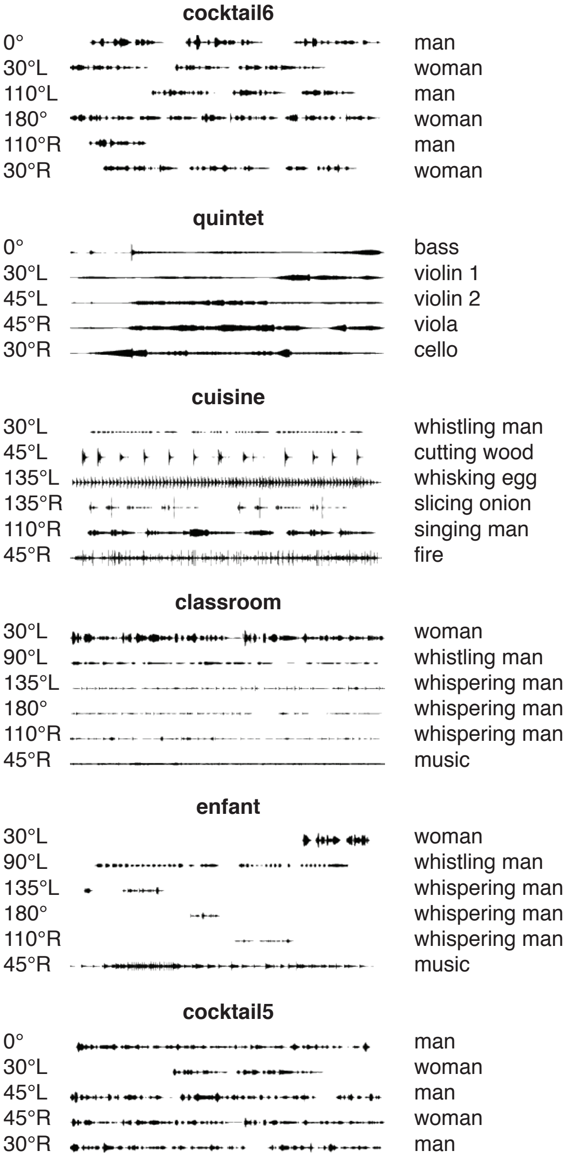

Figure 8: Program material

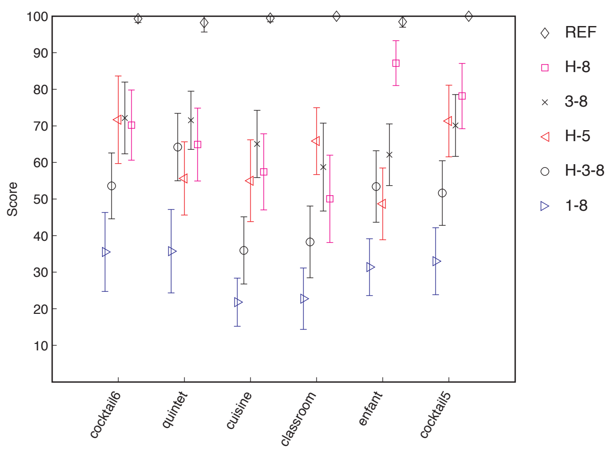

Figure 9: Mean scores across all participants for each system in each scene. Error bars indicate 95% confidence interval.

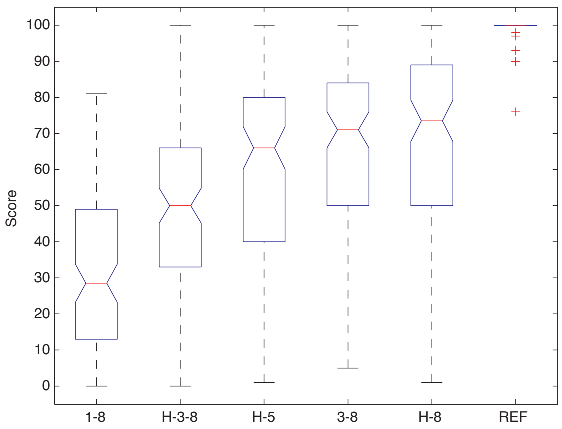

Figure 10: Box plot showing results across all six scenes and nineteen participants. Red lines indicate median scores. The top and bottom of each box are the 25th and 75th percentiles of samples. The width of the notches is calculated so that boxes whose notches do not overlap have different medians at the 5% significance level. Whiskers show minimum and maximum scores, and red crosses indicate outliers.

Nineteen listeners participated in the test. Participants were instructed to concentrate on the perceived direction and spatial sharpness of the sounds, and to ignore any other differences. After receiving instructions, each participant was left to perform the test alone. The test took about 35 minutes.

5.2. Results

The results are summarized in Figures 9 and 10. The difference in score between the H-5, 3-8 and H-8 systems is not statistically significant. The differences between other systems are, however, statistically significant. The difference between the H-5 and H-8 systems is nearly significant, with p = 0.062 in a one-way repeated measures ANOVA test with post-hoc Tukey's HSD [11].

5.3. Discussion

As with any listening test, one may question the generality of the results, given that only a few test sounds can be presented to the listeners. The most likely weakness of non-linear methods is their reproduction of multiple overlapping sound sources, and the scenes that were tested were considerably "busier" than the typical sound scene encountered in broadcasting or telecommunication. The results for the systems using HARPEX should therefore be conservative estimates.

It may also be argued that ambisonics sounds different than pairwise panning, and that the reference signal, being a special case of pairwise panning, biases the test in favor of that panning function. This was the reason for the inclusion of the H-3-8 system, where the panning function is identical to the 3rd order max r_E decoder. The H-3-8 system did indeed score lower than the H-8 system. One may speculate that max r_E decoding and HARPEX each introduce their own differences from the reference signal, and that these accumulate to a greater overall difference in the H-3-8 system. This theory could be tested with a pairwise comparison test.

The most surprising result was perhaps the comparatively high score of the H-5 system. This may be attributed to the reference signals slightly over-using the loudspeakers belonging to the 5.0 subset. This effect is difficult to quantify, since different sound sources will attract different amounts of attention, not only because of differences in amplitude and spectrum, but also because of differences in extrinsic properties. Rigorous comparison of different loudspeaker layouts would require a larger number of loudspeakers, so that the reference could be played over other loudspeakers than either of the systems under test.

Comparing the scores in our experiment to those in [2], one notes several differences: Our 1-8 system scored a mean of 30, while their comparable SoundField system scored 19. The difference is even larger between our 3-8 system, scoring 67, and their 8 sensors system, which scored 31. There are several factors that are likely to contribute to these differences. Firstly, our systems use ideal encoding, while their systems include the measured characteristics of real microphones. Secondly, the decoding in our experiment was to eight loudspeakers, whereas their signals were played over twelve loudspeakers. Thirdly, since their experiment also includes 4th order systems, participants are likely to compress the high end of their scales to accommodate the higher quality of those systems.

Another relevant listening test is described in [12]. Our 1-8 system with mean score 30 should be compared to their A16 and A4 systems, using first-order decoding to 16 and 4 loudspeakers, scoring 34 and 60 points, respectively. The highest-scored system in their test was D16_ideal using the DirAC decoder, at 85 points. This should be compared to H-8 at 68 points. As their experiment included anchors with mono and low-pass filtered sound, it is likely that participants compressed the low end of their scales to accommodate the lower quality of these systems. Also, their D16_ideal system was allowed access the same loudspeakers as the reference signal, thereby avoiding the penalty inherent in panning, incurred by our H-8 system.

5.4. Other observations

In addition to the quantitative results above, the authors have made the following observations in their own listening:

- When comparing the H-8 and 1-8 systems, the H-8 location of the sound is not only closer to the reference but also appears more robust to variations in the listening position. When comparing H-8 and 3-8, although both are similar in terms of proximity to the reference, H-8 appears slightly more robust across a larger sweet spot.

- The 1-8 decoding appears to float ambiguously in the space and is clearly detached from the physical location of the loudspeakers. This is particularly apparent in comparison to H-8. Although recreation of the target is extremely poor, this spatial ambiguity may be useful for certain types of effect which are intended to project a sense of spaciousness or envelopment without needing location accuracy.
- When comparing H-8 and 3-8, the latter is marginally more removed from the physical loudspeakers, and for a single listener central to the sweet spot the result is maybe more pleasing or natural. However, this effect involves a tradeoff in image stability over a larger listening area, where H-8 provides that stability.
- When listening to individual output channels of the H-8 system, one can detect leakage from high-amplitude channels to adjacent, low-amplitude channels. The leakage is frequency-dependent and therefore sounds distorted. When all channels are played on the intended loudspeaker setup, however, the leaking sound is masked by the adjacent original sound and becomes inaudible, even close to the loudspeakers.
- The decoder has four adjustable parameters; window length, time smoothing constant, frequency smoothing constant and minimum angle between planewaves. Each of these constants can be adjusted over a range of about an order of magnitude without audible effect on a selected set of test sounds. When all parameters are set to the middle of their useful range, no artifacts were detected during listening to HRTF decoding of the 208 sound files currently available at the Ambisonia web site [13].
- If the decoder is modified to use only a single direction estimate at each frequency component, the position of some sound sources appear to move around in the auditory scene even though they are presumably immobile. This artifact disappears when two direction estimates are allowed.

6. CONCLUSION

The proposed method provides a means for playing back first-order material over large loudspeaker setups with improved spatial definition and a much larger sweet spot than is possible with the other method that was tested. Surprisingly good results were achieved over a 5.0 setup compared an eight-loudspeaker setup. The artifacts that are audible in a straightforward decoder using HARPEX can be suppressed to safely inaudible levels without giving up noticeable amounts of sharpness in the auditory scene.

7. ACKNOWLEDGMENTS

The authors would like to thank NOTAM for providing studio time for the listening tests, Orange Labs for providing the program material from their experiments, and Øyvind Hammer for advice on the statistical analysis.

8. REFERENCES

[1] M. Gerzon, “Periphony: With-height sound reproduction,” *J. Audio Eng. Soc*, vol. 21, no. 1, pp. 2–10, 1973.

[2] S. Bertet, J. Daniel, E. Parizet, and O. Warusfel, “Influence of Microphone and Loudspeaker Setup on Perceived Higher Order Reproduced Sound Field,” in *1st Ambisonics Symposium*, 2009.

[3] D. McGrath and A. McKeag, “Wavelet conversion of 3-D audio signals,” Sept. 30 2003, u.S. Patent 6,628,787.

[4] V. Pulkki, “Directional audio coding in spatial sound reproduction and stereo upmixing,” in *Proc. of the AES 28th Int. Conf, Pitea, Sweden*, 2006.

[5] C. Faller, “Parametric coding of spatial audio,” Ph.D. dissertation, École Polytechnique Fédérale de Lausanne, 2004.

[6] V. Pulkki, “Virtual sound source positioning using vector base amplitude panning,” *Journal of the Audio Engineering Society*, vol. 45, no. 6, pp. 456–466, 1997.

[7] M. Neukom and J. Schacher, “Ambisonics equivalent panning,” in *Proceedings of the 2008 International Computer Music Conference, Belfast, UK*, 2008.

[8] A. Berkhout, D. De Vries, and P. Vogel, “Acoustic control by wave field synthesis,” *The Journal of the Acoustical Society of America*, vol. 93, p. 2764, 1993.

[9] “Method for the subjective assessment of intermediate quality level of coding systems,” *ITU-R BS*, pp. 1534–1, 2003.

[10] G. Wakefield, “Ambisonics for Max/MSP,” http://www.grahamwakefield.net/soft/ambi~/index.htm, 2006.

[11] Ø. Hammer, D. Harper, and P. Ryan, “PAST: paleontological statistics software package for education and data analysis,” *Palaeontologia electronica*, vol. 4, no. 1, p. 9, 2001.

[12] J. Vilkamo, “Spatial Sound Reproduction with Frequency Band Processing of B-format Audio Signals,” Master’s thesis, Helsinki University of Technology, 2008.

[13] Ambisonia, http://www.ambisonia.com.