DIRECTIONAL CODING OF AUDIO USING A CIRCULAR MICROPHONE ARRAY

Anastasios Alexandridis[⋆†] *Anthony Griffin*[⋆] *Athanasios Mouchtaris*[∗⋆†]

[⋆] FORTH-ICS, Heraklion, Crete, Greece, GR-70013
[†] University of Crete, Department of Computer Science, Heraklion, Crete, Greece, GR-71409

ABSTRACT

We propose a real-time method for coding an acoustic environment based on estimating the Direction-of-Arrival (DOA) and reproducing it using an arbitrary loudspeaker configuration or headphones. We encode the sound field with the use of one audio signal and side-information. The audio signal can be further encoded with an MP3 encoder to reduce the bitrate. We investigate how such coding can affect the spatial impression and sound quality of spatial audio reproduction. Also, we propose a lossless efficient compression scheme for the side-information. Our method is compared with other recently proposed microphone array based methods for directional coding. Listening tests confirm the effectiveness of our method in achieving excellent reconstruction of the sound field while maintaining the sound quality at high levels.

Index Terms— microphone arrays, spatial audio, beamforming

1. INTRODUCTION

Spatial audio systems aim to reproduce a recorded acoustic environment by preserving the spatial information (e.g., [1, 2, 3, 4]). Such systems have applications in the entertainment sector, enabling users to watch movies that feature surround sound or play computer games providing a more immersive gaming experience, etc. In teleconferencing they can facilitate a more natural way of communication.

In this paper we propose a real-time method for coding a sound field at a low bitrate using microphone arrays. Reproduction is possible using a loudspeaker configuration or headphones. The sound field is encoded using one audio signal and side-information. We consider microphone arrays—particularly circular arrays—for spatial audio as they are already used in several applications, such as teleconferencing and providing noise-robust speech capture.

Techniques for coding and reproducing spatial audio, when recording a sound scene, have already been proposed. Directional Audio Coding (DirAC) [5] is based on B-format signals and encodes a sound field using one or more signals along with Direction-of-Arrival (DOA) and diffuseness estimates for each time-frequency element. Versions of DirAC that are based on microphone arrays have also been proposed [6, 7]. In [6] differential microphone array techniques are employed to convert the microphone array signals to B-format. However, a bias in the B-format approximation—as illustrated in [8]—leads to biased estimates that can degrade the spatial impression. The authors of [7] utilize array processing techniques to infer the DOA and diffuseness estimates while the reproduction side remains the same as in [5]. Time-frequency array processing is also used in [9] for binaural reproduction.

The aforementioned methods try to encode the sound field in terms of DOA (and/or diffuseness) estimates for each individual time-frequency element, which requires strong W-disjoint orthogonality (WDO) [10] conditions. WDO assumes that there is only one active source in each time-frequency element, which is not the case when multiple sources are active simultaneously. Moreover, these methods suffer from spatial aliasing above a certain spatial aliasing cutoff frequency which causes erroneous estimates and can degrade the quality of the reconstructed sound field. Our method tries to overcome these problems by employing a per time frame DOA estimation for multiple simultaneous sources (for details see [11, 12, 13]). Based on the estimated DOAs, spatial filtering with a fixed superdirective beamformer separates the source signals that come from different directions. The signals are downmixed into one audio signal that can be encoded with any compression method (e.g, MP3). Each source signal is reproduced according to its estimated DOA. While the source separation part can create musical distortions in the separated signals, all signals are played back together—since our goal is to recreate the overall sound field—which eliminates the musical noise. This is an important result of our work validated by listening tests.

2. PROPOSED METHOD

Our proposed method is divided into the encoding and the reproduction stage. Both stages are real-time, with the encoding stage consuming approximately 50% of the available processing time—including the DOA estimation and coding of the sound field—on a standard PC (Intel 2.53 GHz Core i5, 4 GB RAM). The reproduction stage can also be implemented in real-time since its main operation is amplitude panning (or HRTF filtering for binaural reproduction).

In an anechoic environment where P active sources are in the far-field, the signal recorded at the mth microphone of a microphone array with M sensors is the sum of the attenuated and delayed versions of the individual source signals according to their direction. Note that although the model is simplified, the experiments presented in this paper are performed using signals recorded in reverberant environments. The microphone array signals are transformed into the Short-Time Fourier Transform (STFT) domain. To estimate the number of active sources and their DOAs, we utilize the method of [11, 12, 13], which is capable of estimating the DOAs in real-time and with high accuracy in reverberant environments for multiple simultaneously active sources. The method outputs the estimated number of sources $\hat{P}_k$ and a vector with the estimated DOAs for each source (with 1^o resolution) $\boldsymbol{\theta}_k = \left[\theta_1 \cdots \theta_{\hat{P}_k}\right]$ per time frame k.

The source signals are then separated using a fixed superdirective beamformer. The beamforming process employs $\hat{P}_k$ concurrent beamformers each of them steering its beam to one of the directions in $\boldsymbol{\theta}_k$, resulting in the beamformed signals $B_s(k,\omega)$, $s = 1, \cdots, \hat{P}_k$, with ω being the frequency index. The beamformer filter coefficients are calculated by maximizing the array gain, while maintaining a minimum constraint on the white noise gain [14]:

$$\mathbf{w}(\omega,\theta_s) = \frac{[\epsilon\mathbf{I} + \boldsymbol{\Gamma}(\omega)]^{-1}\,\mathbf{d}(\omega,\theta_s)}{\mathbf{d}(\omega,\theta_s)^H\,[\epsilon\mathbf{I} + \boldsymbol{\Gamma}(\omega)]^{-1}\,\mathbf{d}(\omega,\theta_s)} \tag{1}$$

[∗]This research has been co-financed by the European Union (European Social Fund ESF) and Greek national funds through the Operational Program “Education and Lifelong Learning” of the National Strategic Reference Framework (NSRF) - Research Funding Program: THALES, and the Marie Curie IAPP “AVID-MODE” grant within the 7th European Commission FP.

where $\mathbf{w}(\omega, \theta_s)$ is the $M \times 1$ vector of complex filter coefficients, θ_s is the beamformer's steering direction, $\mathbf{d}(\omega, \theta_s)$ is the steering vector of the array, $\mathbf{\Gamma}(\omega)$ is the $M \times M$ noise coherence matrix (assumed diffuse), $(\cdot)^H$ is the Hermitian transpose operation, $\mathbf{I}$ is the identity matrix, and ϵ is used to control the white noise gain constraint. Fixed beamformers are signal-independent, so they are computationally efficient to implement, facilitating their use in real-time systems, since the filter coefficients for all directions can be estimated offline.

Next, a post-filter is applied to the beamformer output to enhance the source signals. The post-filter constructs $\hat{P}_k$ binary masks. The mask for the sth source is given by [15]:

$$U_s(k,\omega) = \begin{cases} 1, & \text{if } s = \arg\max_p |B_p(k,\omega)|^2, \quad p = 1, \cdots, \hat{P}_k \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

The beamformer outputs are multiplied by their corresponding mask to yield the estimated source signals $\hat{S}_s(k,\omega)$, $s = 1, \cdots, \hat{P}_k$.

Equation (2) implies that for each frequency element only the corresponding element of the source with the highest energy is kept, while the others are set to zero. Thus, the masks are orthogonal, meaning that if $U_s(k,\omega) = 1$ for some frequency index ω and frame index k, then $U_{s'}(k,\omega) = 0$ for $s' \neq s$, which is also the case for the signals $\hat{S}_s$. This observation leads to an efficient encoding scheme for the source signals: we can downmix them to one full spectrum signal by summing them up. Side-information, namely the DOA for each frequency bin, is needed so as the decoder can again separate the source signals. The side-information and the time-domain downmix signal are transmitted to the decoder. An MP3 audio coder can be used to reduce the bitrate. Lossless compression schemes can also be applied to the side-information (Section 3).

Equation (2) can be applied to the whole spectrum or up to a specific *beamformer cutoff frequency*. Spatial audio applications that involve speech signals could tolerate such reduction in the processed spectrum. For the frequencies above the beamformer cutoff frequency, the spectrum from an arbitrary microphone is included in the downmix signal. As there are no DOA estimates available for this frequency range, it is treated as diffuse sound in the decoder and reproduced by all loudspeakers. Incorporating this diffuse part is offered as an optional choice, and we also consider the case where the beamformer cutoff frequency is set to $f_s/2$ (with f_s denoting the sampling frequency), i.e., there is no diffuse part.

In the synthesis stage, the downmix signal is transformed into the STFT domain and, based on the beamformer cutoff frequency, the spectrum is divided into the non-diffuse and diffuse part (if exists). In the case where the downmix signal is encoded with MP3, an MP3 decoder is applied first. For loudspeaker reproduction, the non-diffuse part is synthesized using Vector-Base Amplitude Panning (VBAP) [16] at each frequency element. If a diffuse part is included it is played back from all loudspeakers after scaling by the reciprocal of the square root of the number of loudspeakers to preserve the total energy. For headphone reproduction, each frequency element of the non-diffuse part is filtered with the left and right Head-Related Transfer Functions (HRTFs), according to the DOA assigned to the respective frequency element. The diffuse part (if it exists) is included to both left and right channels after appropriate scaling by $1/\sqrt{2}$ for energy preservation.

3. ENCODING OF SIDE-INFORMATION

Since the DOA estimate for each time-frequency element depends on the binary masks of Equation (2), it is sufficient to encode these masks. The active sources at a given time frame are sorted in descending order according to the number of frequency bins assigned to them. The binary mask of the first (i.e., most dominant) source is inserted to the bitstream. Given the orthogonality property of the binary masks, it follows that we do not need to encode the mask for the sth source at the frequency bins where at least one of the previous $s-1$ masks is one (since the rest of the masks will be zero). These locations can be identified by a simple OR operation between the $s-1$ previous masks. Thus, for the second up to the $(\hat{P}_k - 1)$th mask, only the locations where the previous masks are all zero are inserted to the bitstream. The mask of the last source does not need to be encoded, as it contains ones in the frequency bins that all the previous masks had zeros. A dictionary that associates the sources with their DOAs is also included in the bitstream. For decoding, the mask of the first source is retrieved first. For the mask of the sth source, the next n bits are read from the bitstream, where n is the number of frequencies that all the previous $s-1$ masks are zero. This can be identified by a simple NOR operation.

In this scheme the number of required bits does not increase linearly with the number of sources. On the contrary, for each next source we need less bits than the previous one. It is computationally efficient, since the main operations are simple OR and NOR operations. The resulted bitstream is further compressed with Golomb entropy coding [17] applied on the run-lengths of ones and zeros.

4. RESULTS

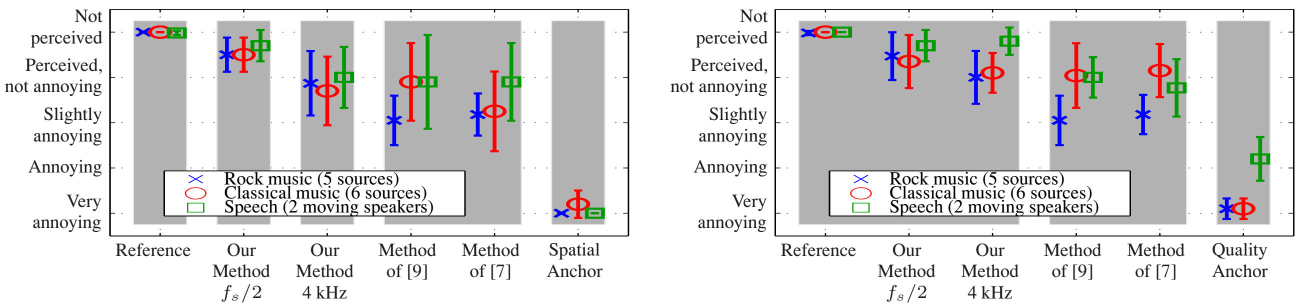
We conducted listening tests on real and simulated microphone array recordings for both loudspeaker and binaural reproduction. We used a uniform circular microphone array with $M = 8$ microphones and a radius $r = 0.05$ m. The sampling frequency was 44.1 kHz. For loudspeaker reproduction we used a circular configuration (radius 1 m) of $L = 8$ uniformly spaced loudspeakers (Genelec 8050) and for binaural reproduction we used high-quality headphones (Sennheiser HD650). The coordinate system used for reproduction places the 0^o in front of the listener, increasing clockwise. The recorded signals were processed using frames of 2048 samples with 50% overlap, windowed with a von Hann window. The FFT size was 4096. Listening tests to test the modelling performance (where the sound scene has been modelled as in Section 2) are presented in Sections 4.1 and 4.2, while results for the modelling with MP3 coding of the downmix signal approach are presented in Section 4.3[1].

4.1. Simulated recordings (modelling performance)

We used the Image-Source Method [18] to produce simulated recordings in a reverberant room of dimensions of $6 \times 4 \times 3$ meters. The walls were characterized by a uniform reflection coefficient of 0.5 and the reverberation time was $T_{60} = 250$ ms. The recordings used were: a 10-second rock music recording with one male singer at 0^o and 4 instruments at 45^o, 90^o, 270^o, and 315^o, which is publicly available from the band "Nine Inch Nails"; a 15-second classical music recording with 6 sources at 30^o, 90^o, 150^o, 210^0, 330^o, and 270^o from [19]; and a 16-second recording with two speakers, one male and one female, starting from 0^o and walking the entire circle at opposite directions. The recordings included impulsive and non-impulsive sounds. Each source was recorded on a separate track and each track was filtered with the estimated Room Impulse Response from its corresponding direction and then added together to form the array recordings.

The listening tests were based on the ITU-R BS.1116 methodology [20]. Ten volunteers participated in each test (authors not included). For the loudspeaker listening test, each track was positioned at its corresponding direction using VBAP (or by filtering it with the corresponding HRTF for the headphone listening test) to create the reference signals. The low-pass filtered (4 kHz cutoff frequency) reference recording served as quality anchor, while the signal at an arbitrary microphone played back from all loudspeakers (or equally from both left and right channels for the headphone listening

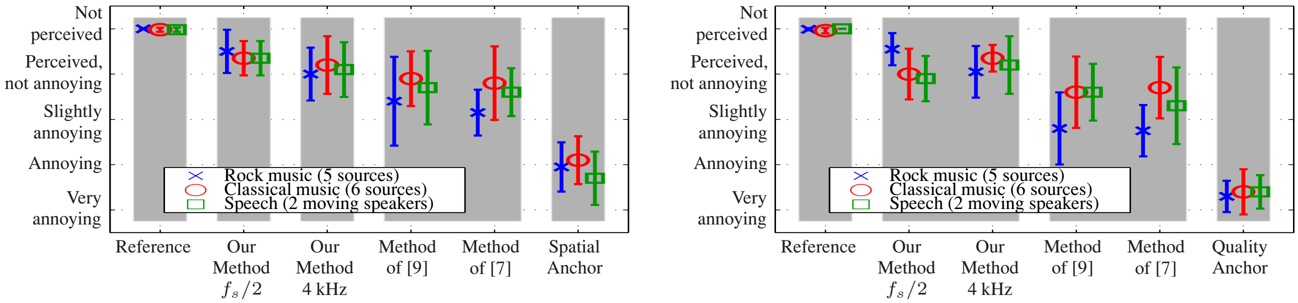
[1] The test samples for our method are available at `http://www.ics.forth.gr/~mouchtar/icassp13_coding.html`



(a) Spatial impression

(b) Sound quality

Fig. 1: Listening test results for simulated recordings with loudspeaker reproduction.



(a) Spatial impression

(b) Sound quality

Fig. 2: Listening test results for simulated recordings with binaural reproduction.

test) was used as a spatial anchor. For HRTF filtering, we used the database of [21]. The subjects (sitting at the "sweet spot" for the loudspeaker test) were asked to compare sample recordings against the reference, using a 5-scale grading. Each test was conducted in two separate sessions: spatial impression and sound quality grading.

Our proposed method with two different beamformer cutoff frequencies, namely, $B = 4$ kHz, and $B = f_s/2$ (no diffuse) was tested against the array-based methods of [9] and [7]. The extension of [9] for loudspeaker reproduction is straightforward by applying VBAP at each frequency element. The DOA estimation method of [7] is based on a linear array, so we used the localization procedure of [9], combining it with the diffuseness and synthesis method of [7].

The mean scores and 95% confidence intervals for the spatial impression and quality sessions for loudspeaker and binaural reproduction are depicted in Figures 1 and 2. An Analysis of Variance (ANOVA) indicates that for both loudspeaker and binaural reproduction a statistical difference between the methods exists in the spatial impression and quality ratings with p-values < 0.01. Multiple comparison tests using Tukey's least significant difference at 90% confidence were performed on the ANOVA results to indicate which methods are significantly different. The methods with statistically insignificant differences have been grouped in gray shading.

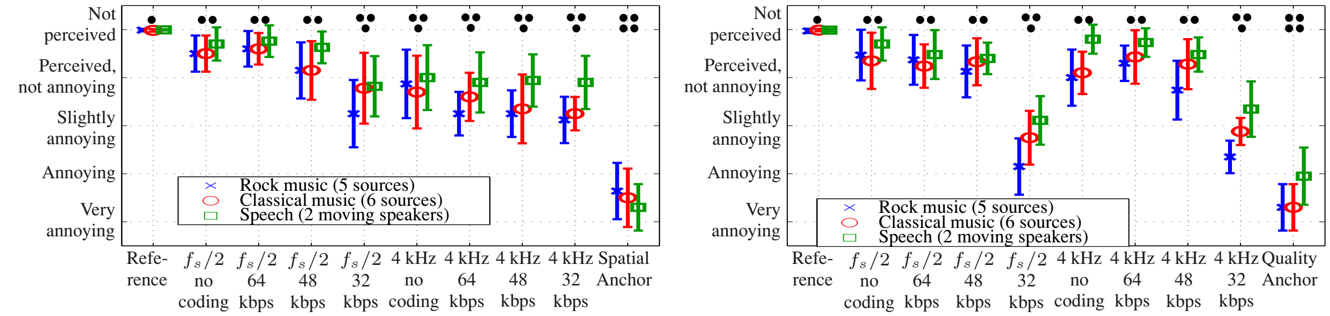
For both types of reproduction, the best results are achieved with our proposed method when $B = f_s/2$ (i.e., no diffuse). With decreasing beamformer cutoff frequency, the spatial impression degrades since directional information is coded only for a limited frequency range. In both versions of our method, the full frequency spectrum is reproduced either from a specific direction or from all loudspeakers (for the diffuse part), so B does not have a severe impact on the sound quality.

Method	Sp. Im.	Q.	Sp. Im.	Q.	Method
Loudspeaker reproduction					
Ours $B = f_s/2$	83%	77%	17%	23%	Method of [9]
Ours $B = f_s/2$	83%	67%	17%	33%	Method of [7]
Ours $B = 4$kHz	63%	67%	37%	33%	Method of [9]
Ours $B = 4$kHz	70%	63%	30%	37%	Method of [7]
Ours $B = f_s/2$	70%	47%	30%	53%	Ours $B = 4$kHz
Method of [9]	67%	33%	33%	67%	Method [7]
Binaural reproduction					
Ours $B = f_s/2$	73%	77%	27%	23%	Method of [9]
Ours $B = f_s/2$	87%	70%	13%	30%	Method of [7]
Ours $B = 4$kHz	57%	63%	43%	37%	Method of [9]
Ours $B = 4$kHz	77%	57%	23%	43%	Method of [7]
Ours$B = f_s/2$	63%	73%	37%	27%	Ours $B = 4$kHz
Method of [9]	77%	57%	23%	43%	Method of [7]

Table 1: Results for the spatial impression (Sp. Im.) and sound quality (Q.) of the preference test. Each row represents a pair of methods with the user preference for each method of a pair.

4.2. Real recordings (modelling performance)

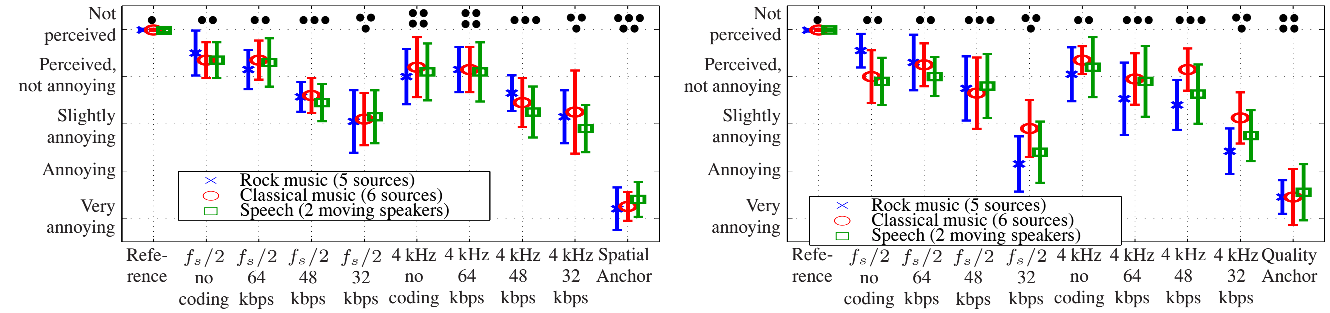
A comparative listening test was conducted with real microphone array recordings. The room dimensions and array specifications were the same as in Section 4.1. We used an array of Shure SM93 omnidirectional microphones and a TASCAM US2000 USB sound card with 8 channels. The recorded test samples were: a 10-second rock music recording with one male singer at 0^o and 4 instruments at 45^o, 90^o, 270^o, and 315^o; a 15-second classical music recording with 4 sources at 0^o, 45^o, 90^o, and 270^o; and a 10-second recording with two male speakers, one stationary at 240^o and one moving clockwise from approximately 320^o to 50^o. Each source signal was reproduced by a loudspeaker (Genelec 8050) located at the corresponding direction at 1.5 m distance. The sound signals were reproduced simultaneously and captured from the microphone array. The music recordings were obtained from the same sources as in the simulated



(a) Spatial impression

(b) Sound quality

Fig. 3: Listening test results with MP3 coding at various bitrates for loudspeaker reproduction



(a) Spatial impression

(b) Sound quality

Fig. 4: Listening test results with MP3 coding at various bitrates for binaural reproduction

	$B = f_s/2$		$B = 4$ kHz	
	Proposed	Huffman	Proposed	Huffman
Rock music	140.57 kbps	166.89 kbps	26.85 kbps	32.55 kbps
Classical music	128.57 kbps	159.20 kbps	22.59 kbps	29.38 kbps
Speech	79.33 kbps	89.10 kbps	12.10 kbps	16.92 kbps

Table 2: Bitrates of the side-information

case. Since a reference recording was not available, we employed a preference test (forced choice). All combinations of our method with $B = f_s/2$ and $B = 4$ kHz and the methods of [9] and [7] were included in pairs and the listeners indicated their preference according to the spatial impression and sound quality in two different sessions. The listening test results for all recordings (Table 1) show a clear preference of our method both in spatial impression and quality.

4.3. Simulated recordings (modelling + coding performance)

To investigate how encoding the downmix audio signal with an MP3 encoder affects the spatial audio reproduction, we conducted a listening test with simulated recordings following the same procedure as in Section 4.1. Our proposed method with $B = f_s/2$ and $B = 4$ kHz and with the mono audio downmix signal encoded at different bitrates, namely 64 kbps, 48 kbps, and 32 kbps, were tested and the subjects were asked to grade the spatial impression and sound quality in two different sessions. The reference and anchor signals were the same as in Section 4.1. We also encoded the side-information using the proposed compression scheme (Section 3). The achieved bitrates for the side-information (with 1^o angle resolution for the DOAs) are shown in Table 2. The Golomb parameter k was set to 2. The bitrates using the Huffman coding on the DOAs are included for comparison. Note that given an angle resolution of 1^o and a 4096-point FFT, the required bitrate for the side-information with no coding is approximately 790 kbps for $B = f_s/2$ which is comparable to the bitrate of an uncompressed audio signal. The bitrates in Table 2 are different for each recording, since the compression depends on the number of sources and the energy contribution of each source. In the classical music no more than 4 sources are simultaneously active, which explains the smaller bitrate compared to the rock music recording which contains 5 simultaneously active sources.

The mean scores and 95% confidence intervals are shown in Figures 3 and 4. A statistical difference exists both in the spatial impression and sound quality ratings for both reproduction types, based on the ANOVA, with p-values < 0.01. To indicate which groups are significantly different, we performed multiple comparison tests using Tukey's least significant difference at 90% confidence. The groups with statistically insignificant differences are denoted with the same symbol at the upper part of Figures 3 and 4. It can be observed that 64 kbps achieves the same results as the modelled uncompressed recording both in spatial impression and quality for both $B = f_s/2$ and $B = 4$ kHz. Noticeable degradation is evident at 32 kbps. The sound quality degradation is more evident in binaural reproduction, since high-quality headphones allow the listeners to notice more easily small quality impairments caused by MP3 coding. In total, our method can utilize a 64 kbps audio signal plus the bitrate for the side-information to encode the sound field without noticeable degradation in the overall quality caused by the coding procedure.

5. CONCLUSIONS

In this paper a real-time method for encoding a sound field using a circular microphone array was proposed. The sound field is encoded using one audio signal and side-information. An efficient compression scheme for the side-information was also proposed. We investigated how coding the audio signal with MP3 affects the spatial audio reproduction through listening tests and found that coding at 64 kbps results in unnoticeable changes compared with the modelled uncompressed case for the same beamformer cutoff frequency. Comparative listening tests with other array-based methods reveal the effectiveness of our method for loudspeaker and binaural reproduction.

6. REFERENCES

[1] J. Breebaart et al., "MPEG Spatial Audio Coding / MPEG Surround: Overview and Current Status," in *119th Audio Engineering Society Convention*, October 2005.

[2] F. Baumgarte and C. Faller, "Binaural cue coding-Part I: Psychoacoustic fundamentals and design principles," *IEEE Transactions on Speech and Audio Processing,*, vol. 11, no. 6, pp. 509 – 519, November 2003.

[3] C. Faller and F. Baumgarte, "Binaural cue coding-Part II: Schemes and applications," *IEEE Transactions on Speech and Audio Processing,*, vol. 11, no. 6, pp. 520 – 531, November 2003.

[4] J. Breebaart, S. van de Par, A. Kohlrausch, and E. Schuijers, "Parametric coding of stereo audio," *EURASIP Journal on Applied Signal Processing*, , no. 1, pp. 1305–1322, 2005.

[5] V. Pulkki, "Spatial sound reproduction with directional audio coding," *Journal of the Audio Engineering Society*, vol. 55, no. 6, pp. 503–516, 2007.

[6] F. Kuech, M. Kallinger, R. Schultz-Amling, G. Del Galdo, J. Ahonen, and V. Pulkki, "Directional audio coding using planar microphone arrays," in *Hands-Free Speech Communication and Microphone Arrays (HSCMA), 2008.*, May 2008, pp. 37–40.

[7] O. Thiergart, M. Kallinger, G. D. Galdo, and F. Kuech, "Parametric spatial sound processing using linear microphone arrays," in *Microelectronic Systems*, Albert Heuberger, Gnter Elst, and Randolf Hanke, Eds., pp. 321–329. Springer Berlin Heidelberg, 2011.

[8] M. Kallinger, F. Kuech, R. Schultz-Amling, G. Del Galdo, J. Ahonen, and V. Pulkki, "Enhanced direction estimation using microphone arrays for directional audio coding," in *Hands-Free Speech Communication and Microphone Arrays (HSCMA), 2008.*, May 2008, pp. 45–48.

[9] M. Cobos, J. J. Lopez, and S. Spors, "A sparsity-based approach to 3D binaural sound synthesis using time-frequency array processing," *EURASIP Journal on Advances in Signal Processing*, vol. 2010, pp. 2:1–2:13, 2010.

[10] S. Rickard and O. Yilmaz, "On the approximate W-disjoint orthogonality of speech," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP) 2002.*, May 2002, vol. 1, pp. 529–532.

[11] D. Pavlidi, M. Puigt, A. Griffin, and A. Mouchtaris, "Real-time multiple sound source localization using a circular microphone array based on single-source confidence measures," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2012*, March 2012, pp. 2625–2628.

[12] D. Pavlidi, A. Griffin, M. Puigt, and A. Mouchtaris, "Source counting in real-time sound source localization using a circular microphone array," in *Sensor Array and Multichannel Signal Processing (SAM 2012)*, Hoboken, NJ, USA, June 17–20, 2012, pp. 529–532.

[13] A. Griffin, D. Pavlidi, M. Puigt, and A. Mouchtaris, "Real-time multiple speaker DOA estimation in a circular microphone array based on matching pursuit," in *European Signal Processing Conference (EUSIPCO 2012)*, Bucharest, Romania, August 27–31, 2012.

[14] H. Cox, R. Zeskind, and M. Owen, "Robust adaptive beamforming," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 35, no. 10, pp. 1365–1376, 1987.

[15] H. K. Maganti, D. Gatica-perez, and I. A. McCowan, "Speech enhancement and recognition in meetings with an audio-visual sensor array," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 8, 2007.

[16] 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.

[17] Solomon W. Golomb, "Run-length encodings," *IEEE Transactions on Information Theory*, vol. 12, no. 3, pp. 399–401, 1966.

[18] E. A. Lehmann and A. M. Johansson, "Diffuse reverberation model for efficient image-source simulation of room impulse responses," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 6, pp. 1429–1439, August 2010.

[19] J. Pätynen, V. Pulkki, and T. Lokki, "Anechoic recording system for symphony orchestra," *Acta Acustica united with Acustica*, vol. 94, no. 6, pp. 856–865, Dec. 2008.

[20] ITU-R, "Methods for the subjective assessment of small impairments in audio systems including multichannel sound systems," 1997.

[21] Gardner B. and K. Martin, "HRTF measurements of a KEMAR dummy-head microphone," in *MIT Media Lab*, May 1994.