# OBJECT-BASED DATA COMPRESSION
# FOR MASSIVE MULTICHANNEL AUDIO

**Maciej BARTKOWIAK, Adam WERESZCZYŃSKI**
Poznań University of Technology, Chair of Multimedia Telecommunications and Microelectronics,
Polanka 3, 60-965 Poznań
mbartkow@multimedia.edu.pl

A new data compression scheme is considered for multichannel audio signal that typically represents a dense sound field. The technique exploits similarities between channels by jointly encoding "objects" extracted from them. For this purpose, dominant energy sources are identified in the spatio-temporal domain, separated from channel data, and encoded jointly, with taking care of appropriate temporal and spectral alignment between channels. The remaining residual signal is encoded in each channel. Experimental results show a significant efficiency increase w.r.t. encoding individual channels without extracting objects. Applications include 3D audio and offline Wave Field Synthesis systems.

## 1. MOTIVATION AND BACKGROUND

Throughout this paper, a massive multichannel audio signal is considered a representation of spatial audio based on a large number (ten or hundreds) of discrete channels. Such representation may be obtained from a bank of densely spaced microphones or rendered through a wave field synthesis system (WFS), and is considered the ultimate spatial audio format which is free from the usual limitations and artifacts of traditional discrete stereophony or multichannel surround sound setups. In the traditional stereo systems, spatial imaging is based on creating illusions of spatial sources based on interaural time and intensity differences. This illusion breaks down with the movement of the listener's head position because the summary wave field created by scarcely distributed loudspeakers does not resemble the spatio-temporal shape of the wave created by the simulated source (fig. 1) and it confuses the auditory system due to inconsistent phase-related spatial cues.

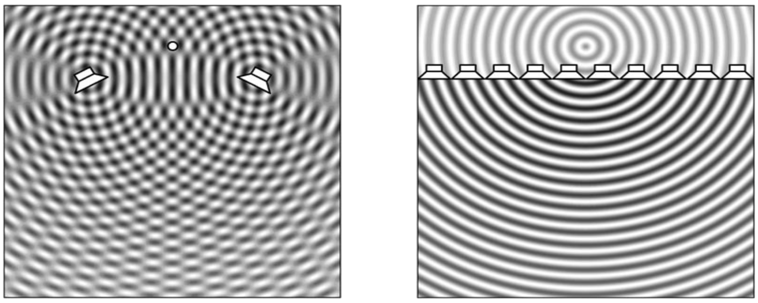

Fig. 1. Left: sound pressure induced by a stereo pair of loudspeakers attempting to reproduce a virtual sound source at the central position, right: sound pressure reproduced by a WFS setup

Wave field synthesis [1,2] offers realistic reproduction of sound pressure distribution within a certain area at the cost of a massive setup of loudspeakers which are densely spaced around the listening area (cf fig. 2). As the reconstruction of the sound field is accurate (up to the frequency of spatial aliasing related to the distance between loudspeakers), the optimal listening position is not limited to the sweet spot, and the effects of acoustical depth and perspective are as much audible as they are in the real environment with physical sound sources located inside the room and behind its walls. Future generations of 3D audio setups related to the free-view television and cinema will certainly employ the WFS principle, as it is already being demonstrated in many commercial systems (e.g. [3]).

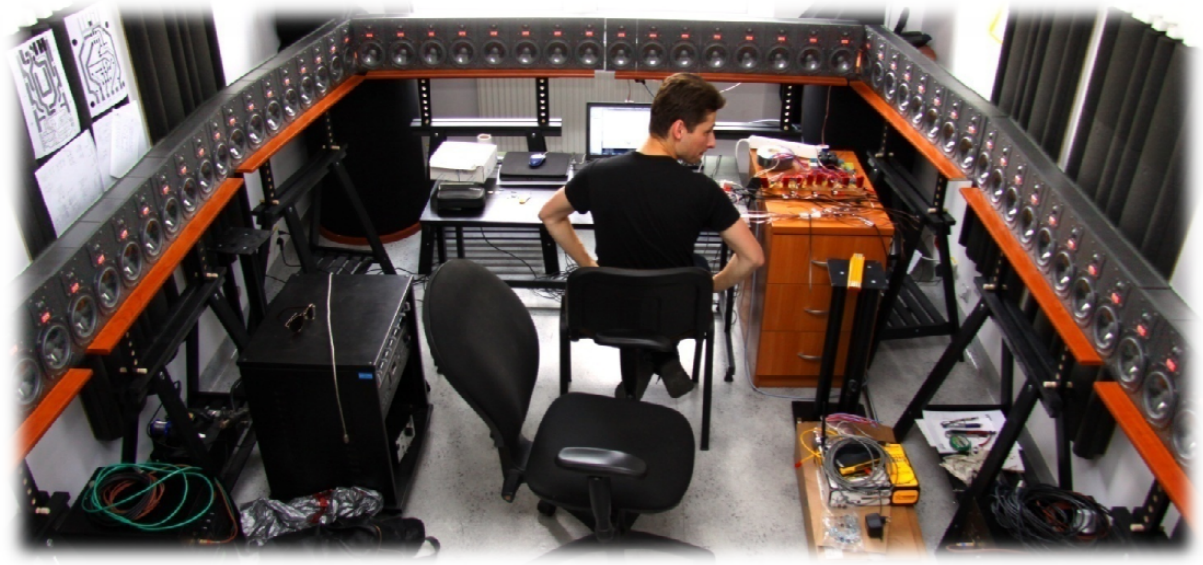

Fig. 2. Experimental WFS system at the Poznań University of Technology

Apart from the prohibitive cost of the WFS physical setup, transmission and digital storage of the massive multichannel audio signal is also a technological challenge. Thus, there is a strong demand for an efficient compression technique that would take into account the nature of the wave field data. A high redundancy may be expected due to the fact that neighbouring channels usually represent close spatial samples of the sound field, hence they transmit very similar content. These similarities may be exploited locally; however, due to spatio-temporal offsets between channel signals increasing with the distance, the correlation cannot be exploited in a global scale via e.g. sum and difference middle-side (MS) coding. Common multichannel (surround sound) systems employ vector-based amplitude panning for positioning virtual sources in the auditory space that ensures mono compatibility which is important from the transmission and reproduction standpoint and makes feasible the matrixing-based compression schemes, for example MPEG-2 [5], or Dolby/ATSC AC3 [6]. This approach cannot be used for WFS data, due to the presence of inter-channel delay.

The Parametric Surround technique, recently developed and included in audio coding standards such as MPEG Surround and SAOC (Spatial Audio Object Coding) [7,8], MPEG USAC (Unified Speech and Audio Coding) [9], as well as the upcoming MPEG 3D Audio [10] is based on parametric representation of the spatial auditory information that is appended to a single mono downmix signal. At the decoder side, individual channel signals are synthesized from the mono signal through manipulations of time delay and amplitude scaling. Such a principle also cannot be used for encoding WFS data, because a downmix of temporally shifted components would yield irreversible comb filtering. Furthermore, parametric technique does not offer transparent quality even at high transmission rates, because it does not preserve the audio waveform. Unfortunately, spatial cues exploited by the WFS are very easily distorted by coding artifacts and would not be properly preserved in parametric coding.

## 2. PROPOSED TECHNIQUE

The main principle of the proposed compression technique (fig. 3) is to decompose the set of M channel signals into K<M individual audio “object” signals and M decorrelated “background” signals, all of which are compressed by a traditional perceptual codec. The assumption is that the object signals correspond to individual sound sources in the space captured by the microphone set, and the background is the residual resulting from the subtraction of the time-aligned object signals from individual channel signals. The number of bits required by an audio codec for high quality representation is very roughly proportional to the energy of the signal. In a perfect scenario, the separation would result in the energy of residual being negligible. Thus, the total number of bits (bulk of which generated by the object signals) would be much lower compared to the total bit stream resulting from individual encoding of each channel. This goal is unrealistic due to imperfect separation, and the background signals usually represent some significant energy, especially in highly reverberant spaces. Experiments show, that even in such cases significant gain in compression efficiency is obtained due to the energy of the background being strongly reduced.

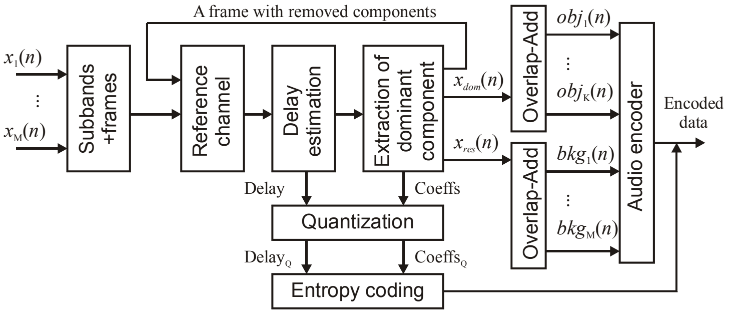

Fig. 3. The general block diagram of the proposed encoder

### 1.1. SOURCE SEPARATION DOMAIN

Classic signal separation techniques known from the literature [11] are unsuccessful in separating sources from a set of mixes involving a significant and time-varying delay. On the other hand, a complete and perfect separation is not necessary for this compression scheme. A novel technique has been developed for this particular task. The multichannel signal is processed in a hybrid spectro-spatio-temporal domain. For this purpose, each channel signal is decomposed into a set of B subbands with a perceptually-motivated perfect reconstruction filterbank. This filterbank employs a simple pyramidal cascade structure of linear phase lowpass filters with decreasing bandwidth, and does not employ subsampling, hence no aliasing artifacts are introduced.

### 1.2. CODEC STRUCTURE

The proposed codec (cf fig. 3) operates on a frame basis. The encoder splits the set of M×B subband signals into short time segments (frames) in order to allow processing sound fields containing moving sources, whose channel contribution changes over time. The frames are overlapping in 50% and windowed for the sake of avoiding any artifacts due to discontinuities, so the original signal may be reconstructed through frame overlap-add (OLA) procedure. The length of frames N=2048 is a compromise between redundancy and the granularity of representation, and corresponds to 23ms at 44.1kHz sampling rate. Subsequently, an iterative greedy procedure attempts at identifying spatio-temporally coherent components which are further combined into object signals. Three steps are essential for estimating a single frame of each individual object signal: channel delay estimation, channel time alignment and principal component analysis. The estimated components are subtracted from each channel signal with appropriate delay and amplitude coefficients, and the process repeats for subsequent components. At each iteration the energy of the residual is measured. Only those components whose subtraction results in a significant decrease of energy are used for the synthesis of the final object signal. The process stops after a pre-defined number of objects is separated. The resulting component and residual frames are subsequently combined in the overlap-add process and form the object and background signals, respectively. These signals are altogether the subject of perceptual audio compression, and the multiplexed encoded data stream is transmitted to the decoder associated by a side information on delays and mixing coefficients.

The decoding procedure is straightforward. The decompressed object as well as decompressed background signals are again decomposed into overlapping frames, wherein each frame of object is added to the corresponding frame of each of the background signals, using an appropriate individual delay and mixing coefficient decoded from the side data. Finally, all frames are combined in the OLA process.

### 1.2. IDENTIFICATION OF THE DOMINANT COMPONENT (SINGLE ITERATION)

In each iteration, a reference channel for the dominant sound source has to be selected. This channel should represent the microphone which is closest to the sound source being currently addressed. Since the iterative procedure in the encoder is greedy, it always selects the source of the highest energy and the best signal-to-noise ratio (here “noise” refers to the mixture of remaining sources, or the residual). For each channel $m$ and subband $b$, the total energy is calculated in the current frame. A two-dimensional energy map $E(m,b)$ is created, roughly representing sound sources that exhibit energy peaks in space and frequency. In a realistic scenario, the spectra of sound sources certainly overlap, which means that most of the subbands contain a mixture of energy from various sources. Since our aim is to identify inter-channel delays, we select only these subbands that offer the highest SNR. A binary mask is used to isolate the dominant components (cf fig. 4). In each subband, only the channel with the highest energy is selected, all remaining channels are rejected. Subsequently, the energy from all remaining subbands is added in each channel. The channel with the highest sum is selected as the reference channel. Thus, all inter-channel delays are calculated w.r.t. this channel. It may happen that the same channel has been previously selected – in such case, all subbands in that channel are zeroed, and the selection procedure repeats.

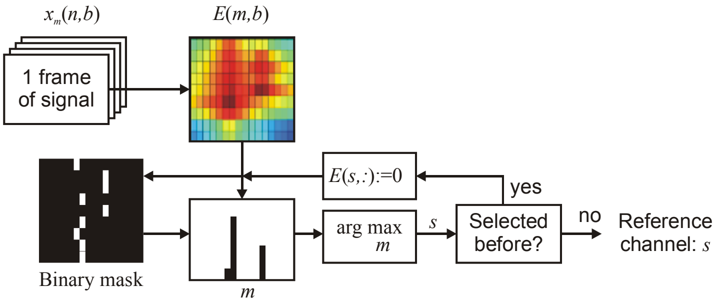

Fig. 4. Selection of the reference channel based on energy maxima in subbands

### 1.3. DELAY ESTIMATION

Knowing the space-frequency location of the dominant component allows to calculate the time delay between corresponding components in other channels w.r.t. selected reference channel $s$. This calculation is based on the estimated inter-channel phase delay (IPD), defined as

$$IPD_s(m,k) = \arg\left\{ X_s(k)\; X_m^*(k) \right\}, \tag{1}$$

where $s$ and $m$ are channel indices, $X_m(k)$ is the Fourier transform of the signal in channel $m$, $k$ is the frequency index, $k$ = 0…N-1, and * denotes complex conjugation. In general, a time delay results in linear phase term in the frequency domain,

$$F\left\{ x_s(n-\Delta n) \right\} = X_s(k)\, \exp\left(-\, j\, 2\pi\, k\, \Delta n\right). \tag{2}$$

However, in a realistic scenario, several important factors impact the values of IPD: presence of noise and signal components from other sources with a different delay, different acoustic conditions for different microphones (e.g. related to sound wave reflections), different frame content resulting from segmentation and windowing. Therefore, the linear trend of the observed IPD (cf fig. 5, left) usually does not correspond to the real delay.

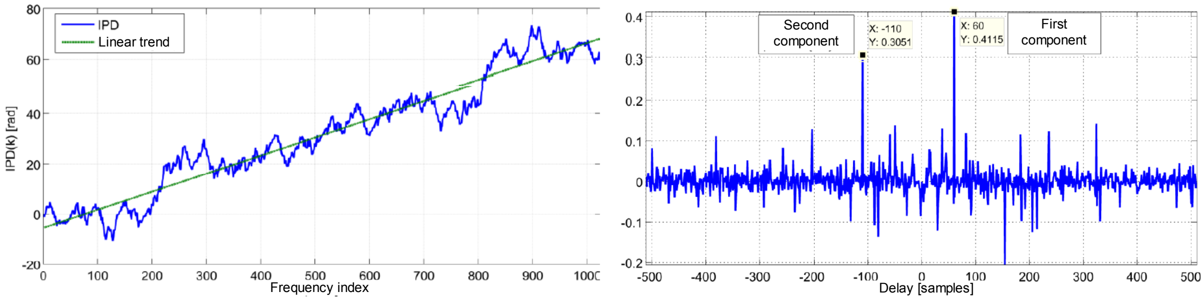

Fig. 5. Left: inter-channel phase delay vs the linear trend; Right: time-domain delay estimation

An original method has been proposed to estimate the delay: the phase difference spectrum is transformed to time domain with a whitened magnitude spectrum,

$$x_{imp}(m,n) = F^{-1}\left\{ \exp\left(-\, j\; IPD_s(m,k)\right) \right\}. \tag{3}$$

This results in a particular time-domain signal $x_{imp}$ that exhibits an impulse at every position corresponding to significant spectral component of distinct group delay (cf fig. 5, right). Experiments show that the location of the largest peak in this signal indicates very reliably the actual time delay between components in channel m w.r.t. channel s. The position of the maximum may be precisely estimated using an interpolation technique (e.g. by fitting a parabolic curve to 3 consecutive samples).

### 1.4. DERIVATION OF OBJECT AND BACKGROUND SIGNALS

Principal Component Analysis (PCA) allows to estimate the dominant common component in a set of data [12]. After the time delay for the dominant component has been estimated in every channel w.r.t. the reference channel, it is possible to compensate all channels in time in such a way that the dominant component is time-aligned, as shown in fig. 6.

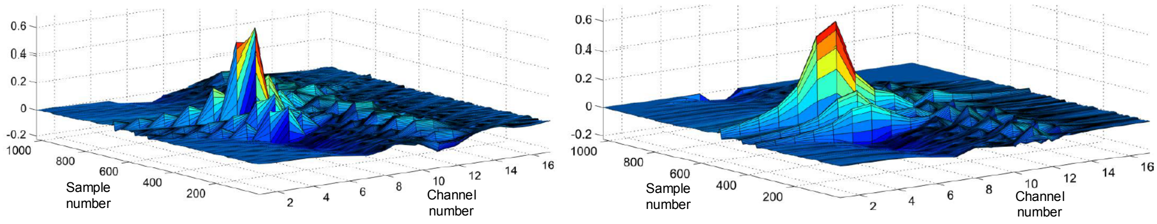

Fig. 6. A single frame of channel signals before and after delay compensation

PCA is performed in time domain for channel signals obtained by mixing all subbands that have not been masked by the binary mask. Vectors V($n$)=[$x_1(n)$, $x_2(n)$, … $x_M(n)$]$^T$ are created from time-aligned samples of all input channels (no windowing is applied within frames), and the covariance matrix $\mathbf{R}_{VV}=\underline{V}\ \underline{V}^T/M$ is calculated for a current frame. Subsequently, a transform matrix **C** is constructed from the eigenvectors of **R** in such a way, that each row of **C** is a solution of

$$\left[\underline{C}_k - \lambda_k \mathbf{I}\right] \underline{X} = 0, \tag{3}$$

where $\lambda_k$ is the $k$-th eigenvalue of **R**, **I** is the identity matrix, and $\underline{X}$ = [1, 1, … 1]$^T$. The first eigenvector $\underline{C}_1$ associated with the largest eigenvalue $\lambda_{\max}$ allows to project the data in all channels onto the axis of largest variance in M-dimensional space

$$x_{dom}(n) = \underline{C}_1\ \underline{V}(n), \tag{3}$$

and results in one frame of the dominant component, $x_{dom}$. Combining all frames in the OLA process results in the object signal of the current iteration of separation. In order to derive the vector of residual signals $\underline{V}_{res}$ for subsequent iterations, the $x_{dom}$ signal needs to be subtracted from all channel signals $\underline{V}$ using appropriate weights,

$$\underline{V}_{res}(n) = \underline{V}(n) - \underline{D}_1 x_{dom}(n)\,. \tag{3}$$

Here, $\underline{D}_1$ denotes a column vector being the first column of $\mathbf{D} = \mathbf{C}^{-1}$, which may be obtained through a least-squares pseudo-inverse. The residual signals, after reverting the time alignment, are combined together for subsequent iterations of the separation algorithm.

### 1.5. DATA CODING

The set of object signals and all background signals need to be encoded by the core codec. Any particular lossy or lossless technique may be used here, for example a multichannel perceptual codec compliant with one of the MPEG standards, provided the bit rate is high enough to assure transparent quality of the reconstructed audio. Care should be taken of the proper calculation of masking thresholds in order to avoid unmasking the quantization noise in the process of mixing objects and backrounds at the decoder side. Therefore, for each background signal, a masking profile of the original corresponding channel signal should also be estimated, and the minimum of the two should be used for quantization control.

An additional control problem is related to a proper bit allocation between the object signals and the background signals given target total bit rate. In fact, the optimal division depends on the complexity of the data, the number of channels and the number of object signals. In most cases, the dependence between the bit rate (and quality) of background signals and the quality of reconstructed channels signals is much weaker compared to the object signals (cf. fig 7). Therefore the general rule should be to allocate a large part of the target bit stream to object signals.

The side information containing frame delays and mixing coefficients (the elements of the $\underline{D}_1$ vector) are encoded using a vector quantization technique. The total data stream generated is 96 bits per frame on average, which is insignificant compared to encoded audio data.

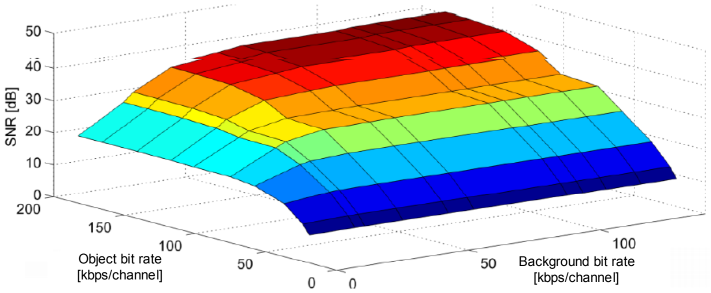

Fig. 7. Dependence between bit rates of object and background signals and the reconstruction quality (SNR)

## 2. EXPERIMENTAL RESULTS AND DISCUSSION

The proposed technique has been extensively tested in various scenarios involving different number of physical sources (fixed and moving) represented by 16 up to 64 channels of audio. Also, several variants of subband decomposition and frame arrangements have been investigated. A standard MPEG-4 AAC technique has been used as the core codec. In all experiments, the same codec has been also employed for direct encoding of the channel signals and served as a reference. Due to the number of experiments, the results (fig 8 and 9) are presented as SNR figures vs the total bit rate for all channels only. The subjective evaluation has been done only in particular cases with best objective results.

Two important classes of experiments show that the overall efficiency of the proposed technique strongly depends on the complexity of the signal. In case of a low number (e.g. 2) of sound sources being active in every time instant, we observe a significant gain in compression efficiency compared to the reference, as shown in fig. 8.

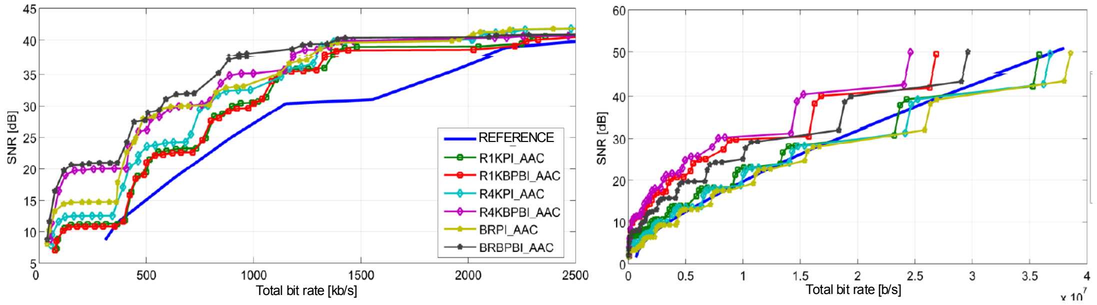

Fig. 8. Reconstructed signal quality vs total bitrate for two fixed (left diagram) and two moving sources (right diagram). The stepping character of curves is a result of varying bitrates for background vs varying bitrates for object signals

Unfortunately, as the complexity of the scene increases, there is a considerable drop in efficiency, up to the point where object-based approach is even less efficient than independent coding of channels (cf fig. 9). Since the performance diagrams for such cases were almost unreadable, we show a difference in SNR w.r.t. the reference technique. As it can be easily observed, only certain combinations of bit allocation (object vs background) result in a positive ΔSNR.

Subjective evaluation of the experimental results has been done in the form of informal listening tests. The conclusions from the tests generally confirm the SNR figures. Transparent audio quality has been obtained at bit rates changing in a wide range, depending on the complexity of the recording (cf tab. 1).

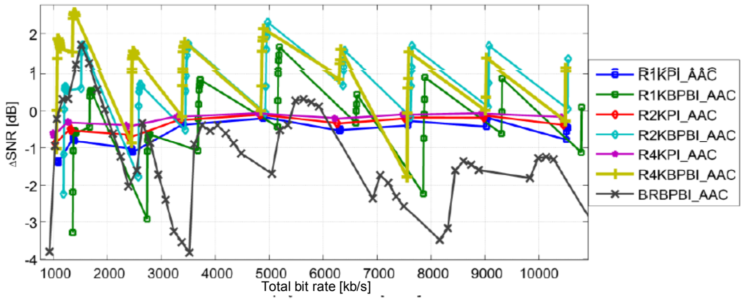

Fig. 9. Compression efficiency compared to the reference technique (positive ΔSNR means advantage over independent compression) for varying bit rate: 16 channels, 6 sound sources active simultaneously.

Table 1. Minimum total bit rate for transparent quality [kb/s]

| Number of sources | **Proposed** technique | **Reference** technique |
|---|---|---|
| 1 fixed | 130-202 | |
| 1 moving | 1319-3790 | 2400-3592 |
| 2 fixed | 673-2637 | 1929-2986 |
| 6 fixed | 4223-8700 | 4722-11740 |

The rather obvious conclusion from these experiments is that the decreased performance for complex mixes is related to more challenging the problem of source separation. Future research will concentrate on improving the separation method.

## REFERENCES

[1] Spors S., Teutsch H., Rabensteing R, *High-Quality Acoustic Rendering with Wave Field Synthesis.* In: Proc. of the Vision, Modeling, and Visualization Conf. 2002 (VMV 2002), Erlangen, 2002, 101-108.

[2] Verheijen E., *Sound Reproduction by Wave Field Synthesis.* Doctoral dissertation, Technische Universitaet Delft, 2010.

[3] Rébillat, M., Katz, B. F., Corteel, E, *SMART-I 2: Spatial multi-user audio-visual real-time interactive interface*, *A broadcast application context*. In: 3DTV Conference: The True Vision-Capture, Transmission and Display of 3D Video, IEEE, 2009, 1-4.

[4] Schleusing O., *Sonic Emotion Technology Presentation*, Sonic Emotion A.G., 2013

[5] ISO/IEC JTC1/SC29/WG11 MPEG, *International Standard ISO/IEC 13818-3, Generic Coding of Moving Pictures and Associated Audio: Audio*, Int. Org. Stand., 1997

[6] Advanced Television Systems Committee, *Digital Audio Compression Standard (AC-3, E-AC-3)*. Document: A/52:2010, ATSC Inc, Washington D.C., 2010

[7] ISO/IEC MPEG, *International Standard ISO/IEC 23003-1: Information Technology – Part 1: MPEG Surround*. Int. Org. Stand., 2007

[8] ISO/IEC MPEG, *International Standard ISO/IEC 23003-2: Information Technology – Part 2: MPEG Spatial Audio Object Coding*. Int. Org. Stand., 2010

[9] ISO/IEC MPEG, *International Standard ISO/IEC 23003-3: Information Technology – Part 3: MPEG Unified Speech and Audio Coding*. Int. Org. Stand., 2012

[10] ISO/IEC MPEG, *International Standard ISO/IEC 23008-3: Information technology — High efficiency coding and media delivery in heterogeneous environments — Part 3: 3D audio*. Int. Org. Stand., 2014

[11] Comon P., Jutten Ch. (eds), *Handbook of Blind Source Separation: Independent Component Analysis and Applications*. Academic Press, 2010

[12] Jolliffe I.T., *Principal Component Analysis (Springer Series in Statistics)*. Springer, 2002