Multichannel Object-Based Audio Coding with Controllable Quality

Stanislaw Gorlow, Emanuël A. P. Habets, Sylvain Marchand

MULTICHANNEL OBJECT-BASED AUDIO CODING WITH CONTROLLABLE QUALITY

Stanislaw Gorlow, Emanuël A. P. Habets† and Sylvain Marchand‡*

*Univ. Bordeaux, LaBRI, UMR 5800, 33400 Talence, France
†International Audio Laboratories Erlangen, 91058 Erlangen, Germany
‡Univ. Brest, Lab-STICC — CNRS, UMR 6285, 29238 Brest, France

ABSTRACT

In this paper a new multichannel object-based audio coding scheme with scalable signal quality is proposed. The novel scheme is based on controlled downmixing and demixing. By means of a dedicated control mechanism, a number of distinct audio objects are mixed into a lower number of channels. The latter is chosen such that the desired quality level is met after demixing. The quality is assessed with two new psychoacoustically motivated metrics. Following the informed source separation approach, the downmix is decomposed via optimum spatial filtering guided by short-time power spectral densities of the audio objects. In an experiment it is shown that the raw data rate of an exemplary 10-track recording can be reduced by at least 30 % using linear pulse-code modulation while maintaining perceptual transparency.

***Index Terms*—** Audio coding, multichannel, object-based, quality control, spatial filtering

1. INTRODUCTION

Ever since the end of the last century, coding of audiovisual objects has been of particular interest to the Moving Picture Experts Group (MPEG), and it has gained importance in recent years. Whereas the first audio coders were all channel-based, a paradigm shift towards source-based coding was initiated by works like [1]. A more recent example is MPEG's Spatial Audio Object Coding (SAOC) [2–4] or the work in [5]. The necessity for object-based coding in the sense of sound sources arises when distinct audio objects are to be stored or transmitted for the purpose of post-hoc reproduction in different environments. So far, its application fields include remixing, video gaming, home cinema or 3D audio, and there might be more in the future.

The work presented here focuses on the question how a number of given source signals or objects can be represented by a reduced number of mixture channels and recovered using the mixture and a small amount of metadata. The work by Faller [1] considers only single-channel mixtures and has no means to scale the quality after demixing. The resulting quality can so be expected to be the worst possible, as a single-channel mixture exhibits the highest overlap between objects. Whereas in [5] Hotho *et al.* generalize the mixture to more than one channel and propose to use the residual to scale the audio quality up to perceptual transparency, there is no explicit control over the quality, except that the latter is said to improve with the bandwidth of the residual signal by rule of thumb. Moreover, the works in [1] and [5] evaluate the quality empirically after rendering the decoded objects into a prescribed format such as 5.1 surround and are consequently bound to the sound reproduction system.

Though related to previous approaches, our work capitalizes on quality-driven demixing which is further independent of mixing and rendering after demixing. Moreover, we pursue the informed source separation approach [6]. It has been recently demonstrated in [7, 8] that an underdetermined linear mixture can be decomposed into an arbitrary number of components by means of spatial filtering. When the separation is carried out in the short-time Fourier transform or STFT domain, the estimates show distortion in amplitude and phase. It is clear that the amount of distortion, which is due to bleed from other sources but also the filter response, decreases with the number of mixture channels because the separation problem becomes better conditioned and the array gain increases. In this work we show how these facts can be exploited to code audio objects in a controllable way.

The organization of the paper is as follows. The signal model and the problem to solve are stated in Section 2. Section 3 outlines the proposed coding scheme, thereby focusing on the mixdown and the demix. Section 4 introduces the quality metrics and the control mechanism. The proposed scheme is tested on a 10-track recording in Section 5. Section 6 concludes the paper and mentions directions for future work.

2. SIGNAL MODEL AND PROBLEM STATEMENT

Using the STFT signal representation, in each frequency subband k, the source signals $\{s_{ik}(n)\}$, $i = 1, 2, \ldots, I$, are circular symmetric complex normal stochastic processes with zero mean and a diagonal covariance matrix $\mathbf{R}_{\mathbf{s}_k}(n) = \operatorname{diag}\left[\phi_{s_{1k}}(n), \phi_{s_{2k}}(n), \ldots, \phi_{s_{Ik}}(n)\right]$ that evolve over discrete time n, where $\{\phi_{s_{ik}}(n)\}_k$ is moreover the short-time power spectral density (STPSD) of the ith source signal. The source signals are linearly combined into an M-channel mixture signal ($M < I$) according to

$$\mathbf{x}_k(n) = \sum_{i=1}^{I} \mathbf{a}_i s_{ik}(n) = \mathbf{A}\mathbf{s}_k(n), \tag{1}$$

where $\mathbf{A} \in \mathbb{R}^{M \times I}$ represents an instantaneous mixing system that is assumed real for practical reasons. Our objective is formulated in the following manner. Given the mixing rule[1] and the STPSDs of the source signals $\left\{\boldsymbol{\phi}_{\mathbf{s}_k}(n) = \left[\,\phi_{s_{1k}}(n)\ \phi_{s_{2k}}(n)\ \cdots\ \phi_{s_{Ik}}(n)\,\right]\right\}_k$, find a low-rank signal space representation $\{\mathbf{x}_k(n)\}_k$ which satisfies a minimum-similarity constraint on the recovered source signals after

This research was partially funded by the "Agence Nationale de la Recherche" within the scope of the DReaM project (ANR-09-CORD-006).

The International Audio Laboratories Erlangen is a joint institution of the Fraunhofer IIS and the University of Erlangen-Nuremberg.

[1] The term "mixing rule" means a set of distinct relations between input and output variables including the mixing system but also its definition.

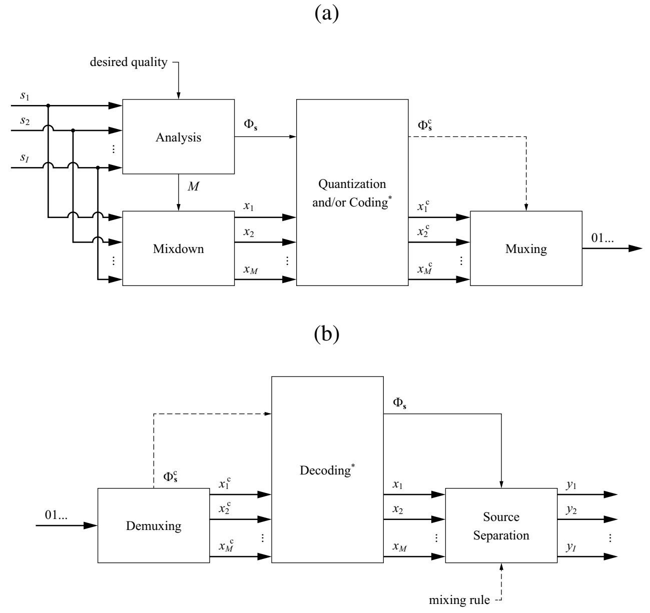

Fig. 1. Functional block diagrams of (a) the encoder and (b) the decoder. The asterisk indicates that the coding/decoding blocks may also include watermarking/signing functionality.

transformation back to the original signal space. Or in other words, what is the minimum number of channels $M_{\min}$ into which one can mix the source signals and yet maintain a desired quality level after demixing. The quality metric shall further relate to, but not model, human perception.

3. PROPOSED CODING SCHEME

3.1. System overview

The proposed coding scheme comprises an encoder and a decoder. Its functional principle is depicted in the form of a block diagram in Fig. 1. The analysis block performs the computation of the STPSDs of all I source signals, as indicated by $\mathbf{\Phi_s}$. From $\mathbf{\Phi_s}$, the number of required mixture channels M is derived that guarantees the desired quality on the decoder side. This is accomplished through a quality control mechanism that is discussed in Section 4. The STPSDs are quantized on an ERB-log frequency-power scale and, for example, differential-entropy coded [8]. In addition, the Free Lossless Audio Codec (FLAC) [9] can be used to reduce the file size of the mixture signal. When FLAC is the coder of choice, which is “lossless”, the STPSDs can be attached to the mixture in the form of a watermark [10, 11] before coding. Otherwise, they are embedded into a serial bitstream as a supplement to the encoded audio data. In the decoder, the demultiplexer reassembles the encoded mixture signal from the bitstream and the metadata if necessary. If lossless compression is used in the encoder, the decoding block may as well be followed by watermark extraction. The decoded STPSDs accompany the source separation that is discussed in more detail in the follow-up sections together with the mixdown.

3.2. Mixdown

Due to the fact that we can decide freely about the mixing rule, we will seek to implement the mixdown such that the decomposition of the mixture in the decoder becomes *controllable* and in this way the resulting signal quality predictable. It is also highly desirable for the signal quality *not* to depend on the mixing rule but on the number of mixture channels only. To accomplish this, one must consider how the decomposition is carried out.

3.2.1. Optimum filters and low-order statistics

A spatial filter that maximizes the signal-to-interference ratio in the mean-square error (MSE) sense has the generic form [12]

$$\mathbf{w}_{ik\mathrm{o}}(n) = \alpha_{ik}(n)\mathbf{R}_{\mathbf{x}_k}^{-1}(n)\mathbf{a}_i, \tag{2}$$

$\alpha \in \mathbb{C}$, where $\mathbf{R_x}$ is a nonsingular mixture covariance matrix and $\mathbf{R_x}^{-1}$ is its inverse. If the mixing matrix $\mathbf{A}$ and the input covariance matrix $\mathbf{R_s}$ are known, $\mathbf{R_x}$ is also known, since

$$\mathbf{R}_{\mathbf{x}_k}(n) = \mathbf{A}\mathbf{R}_{\mathbf{s}_k}(n)\mathbf{A}^\mathsf{T}, \tag{3}$$

where superscript T denotes the transpose. Here in our case, $\mathbf{R_x}$ is real symmetric and as such positive-semidefinite. Specific problems may require the filter response to be constrained in order to obtain a better suited solution. And so, α is formulated differently from one filter to another. One well-known example is the minimum-variance distortionless response (MVDR) filter [13] which has a unity-gain response with zero phase shift. The corresponding weight vector is

$$\mathbf{w}_{ik\mathrm{o}}^{\mathrm{MVDR}}(n) = \frac{\mathbf{R}_{\mathbf{x}_k}^{-1}(n)\mathbf{a}_i}{\mathbf{a}_i^\mathsf{T}\mathbf{R}_{\mathbf{x}_k}^{-1}(n)\mathbf{a}_i}. \tag{4}$$

The distortionless response property of the MVDR filter is used in Section 4 to define a similarity metric.

3.2.2. Signal-to-interference ratio and array gain

An estimate for the ith source component in the kth frequency bin and the nth time segment is given by

$$y_{ik}(n) = \mathbf{w}_{ik}^\mathsf{H}(n)\mathbf{x}_k(n), \tag{5}$$

where superscript H denotes Hermitian or conjugate transpose. The corresponding STPSD value is

$$\phi_{y_{ik}}(n) = \mathrm{E}\left[|y_{ik}(n)|^2\right] = \mathbf{w}_{ik}^\mathsf{H}(n)\mathbf{R}_{\mathbf{x}_k}(n)\mathbf{w}_{ik}(n), \tag{6}$$

where E denotes expectation. Using (3), (6) can also be written as

$$\begin{aligned}\phi_{y_{ik}}(n) = {} & \underbrace{\left|\mathbf{w}_{ik}^\mathsf{H}(n)\mathbf{a}_i\right|^2 \phi_{s_{ik}}(n)}_{\text{signal of interest}} \\ & + \underbrace{\sum_{l=1,l\neq i}^{I} \left|\mathbf{w}_{ik}^\mathsf{H}(n)\mathbf{a}_l\right|^2 \phi_{s_{lk}}(n)}_{\triangleq\, \phi_{b_{ik}}(n),\ \text{residual interference or bleed}} .\end{aligned} \tag{7}$$

The output signal-to-interference ratio is then

$$\begin{aligned}\mathrm{SIR}_{ik}^{\mathrm{out}}(n) = {} & \underbrace{\frac{\phi_{s_{ik}}(n)}{\sum_{p=1,p\neq i}^{I}\phi_{s_{pk}}(n)}}_{\triangleq \mathrm{SIR}_{ik}^{\mathrm{in}}(n)} \\ & \cdot \underbrace{\frac{\left|\mathbf{w}_{ik}^\mathsf{H}(n)\mathbf{a}_i\right|^2 \sum_{p=1,p\neq i}^{I}\phi_{s_{pk}}(n)}{\sum_{q=1,q\neq i}^{I}\left|\mathbf{w}_{ik}^\mathsf{H}(n)\mathbf{a}_q\right|^2 \phi_{s_{qk}}(n)}}_{\triangleq G_{ik}(n) > 1},\end{aligned} \tag{8}$$

where $\mathrm{SIR}^{\mathrm{in}}$ is the input signal-to-interference ratio and G is the array gain. The array gain can be shown to be

$$G_{ik}(n) = \frac{\left[\mathbf{a}_i^\mathsf{T}\mathbf{R}_{\mathbf{x}_k}^{-1}(n)\mathbf{a}_i\right]^2 \sum_{p=1,p\neq i}^{I} \phi_{s_{pk}}(n)}{\sum_{q=1,q\neq i}^{I} \left[\mathbf{a}_i^\mathsf{T}\mathbf{R}_{\mathbf{x}_k}^{-1}(n)\mathbf{a}_q\right]^2 \phi_{s_{qk}}(n)} \tag{9}$$

for real $\mathbf{A}$ and real or complex α. As can be seen from (9), the array gain is a function of the mixing system and the STPSDs.

3.2.3. Mixing system

The mixing system is designed as an M-element vertical line array and the ith source is associated with an angle θ_i,

$$\theta_i = \frac{\pi}{I+1} \cdot i, \tag{10}$$

for $i = 1, 2, \ldots, I$. θ can be thought of as the angle between the propagation path and the normal to the array axis in a 2D, i.e. two-dimensional, sound field. The mixing coefficients are calculated as

$$a_{m+1,i} = \cos{(m\theta_i)}, \tag{11}$$

for $m = 0, 1, \ldots, M-1$, where $\cos{(m\theta)}$ represents an mth-order Chebyshev polynomial in $\cos{(\theta)}$. As a consequence of (10), $\mathbf{A}$ has linearly independent columns and because of (11), $\mathbf{A}$ is real and has full row rank. Nonetheless, $\|\mathbf{a}_i\| \neq \|\mathbf{a}_l\|$ if $i \neq l$.[2]

As previously stated, it is highly desirable that the quality of the estimates is independent of the mixing rule. It is hence vital to make sure that the output signal-to-interference ratio in (8) is the same for all sources. This is accomplished with Algorithm 1 which under the assumption that all I mixture components are standard normal, and with knowledge of the mixing rule, provides the input variances that yield an equal output signal-to-interference ratio for all sources. In this way, one compensates for differences in "radiation" patterns.[3]

3.3. Source separation

Equations (1), (10) and (11) constitute the mixing rule which is used on the encoder side during mixdown. Having knowledge of this rule on the decoder side means knowing the mixing matrix $\mathbf{A}$, provided that the number of objects I is known, too. The transmission of the mixing coefficients can hence be omitted. Using (2), (3) and (5), we can formulate a joint demixing operation according to

$$\mathbf{y}_k(n) = \mathbf{W}_k^\mathsf{T}(n)\mathbf{x}_k(n). \tag{12}$$

Moreover, as the local constellation of mixture components changes with time and frequency, we distinguish between *inactive* and *active* time-frequency (TF) points (k, n). Active points can be determined, overdetermined or underdetermined. The number of components in a TF point, denoted as $I_k(n)$, and also their indices can be inferred from the signaled STPSDs $\left\{\boldsymbol{\phi}_{\mathbf{s}_k}(n)\right\}_k$. Taking all this into account, the demixing matrix $\mathbf{W}^\mathsf{T}$ for an active TF point (k, n) is given by

$$\mathbf{W}_k^\mathsf{T}(n) = \begin{cases} \mathbf{A}^{-1} & \text{if } I_k(n) = M \\ \mathbf{A}^{+} & \text{if } I_k(n) < M \\ \operatorname{diag}\left[\alpha_{ik}(n)\right]_{i=1}^{I_k(n)} \mathbf{A}^\mathsf{T}\mathbf{R}_{\mathbf{x}_k}^{-1} & \text{if } I_k(n) > M \end{cases} \tag{13}$$

$\forall(k, n)$, where $\mathbf{A}^{+}$ is the Moore–Penrose pseudoinverse and

$$\alpha_{ik}(n) = \sqrt{\frac{\phi_{s_{ik}}(n)}{\mathbf{a}_i^\mathsf{T}\mathbf{R}_{\mathbf{x}_k}^{-1}(n)\mathbf{a}_i}} \tag{14}$$

is the weight of the *power-conserving minimum-variance* (PCMV) filter [8]. As $\alpha \in \mathbb{R}$ in (14), the phase response of the filter is free of distortion. Equation (14) can also be derived by plugging (2) into (6) and solving (6) for $|\alpha_{ik}|$ so that $\phi_{y_{ik}} = \phi_{s_{ik}}$.

Algorithm 1 Equal-$\mathrm{SIR}^{\mathrm{out}}$ power distribution

```
function PowDist($I, M, \epsilon$)
    for $i \leftarrow 1, I$ do
        $\theta_i \leftarrow \pi/(I+1) \cdot i$
        for $m \leftarrow 0, M-1$ do
            $a_{m+1,i} \leftarrow \cos{(m\theta_i)}$
        end for
        $\phi_{b_i} \leftarrow 1$
    end for
    $oldcost \leftarrow 0$
    $cost \leftarrow \infty$
    while $|cost - oldcost| > \epsilon$ do
        for $i \leftarrow 1, I$ do
            $\phi_{s_i} \leftarrow \phi_{b_i} / \sum_{l=1}^{I} \phi_{b_l}$
        end for
        $\mathbf{R}_\mathbf{x} \leftarrow \sum_{i=1}^{I} \phi_{s_i}\mathbf{a}_i\mathbf{a}_i^\mathsf{T}$
        $oldcost \leftarrow cost$
        $cost \leftarrow 0$
        for $i \leftarrow 1, I$ do
            $\phi_{y_i} \leftarrow 1/\left(\mathbf{a}_i^\mathsf{T}\mathbf{R}_\mathbf{x}^{-1}\mathbf{a}_i\right)$
            $\phi_{b_i} \leftarrow \phi_{y_i} - \phi_{s_i}$
            $\mathrm{SIR}_i^{\mathrm{out}} \leftarrow \phi_{s_i}/\phi_{b_i}$
            $l \leftarrow \max{(i-1, 1)}$
            $cost \leftarrow cost + \left|\mathrm{SIR}_i^{\mathrm{out}} - \mathrm{SIR}_l^{\mathrm{out}}\right|$
        end for
    end while
    return $(\phi_{s_1}, \phi_{s_2}, \ldots, \phi_{s_I})$
end function
```

4. QUALITY CONTROL MECHANISM

4.1. Quality metrics

We define a similarity index (SIX) according to

$$\begin{aligned} \mathrm{SIX}_{iz}(n) = &\left\{1 - \min\left[\frac{|\phi_{y_{iz}}(n) - \phi_{s_{iz}}(n)|}{\phi_{s_{iz}}(n)}, 1\right]\right\} \\ &\cdot \frac{\phi_{y_{iz}}(n) - \phi_{b_{iz}}(n)}{\phi_{y_{iz}}(n)}, \end{aligned} \tag{15}$$

$\mathrm{SIX} \in [0, 1]$, where z is the band index on an ERB-like frequency scale, see [8]. For the PCMV filter, (15) simplifies to

$$\mathrm{SIX}_{iz}^{\mathrm{PCMV}}(n) = \frac{\phi_{s_{iz}}(n)}{\phi_{y_{iz}}^{\mathrm{MVDR}}(n)} = \phi_{s_{iz}}(n)\mathbf{a}_i^\mathsf{T}\mathbf{R}_{\mathbf{x}_z}^{-1}(n)\mathbf{a}_i. \tag{16}$$

The relation between $\mathrm{SIR}^{\mathrm{out}}$ and $\mathrm{SIX}^{\mathrm{PCMV}}$ is given by

$$\begin{aligned} \mathrm{SIR}_{iz}^{\mathrm{out}}(n) &= \frac{\phi_{s_{iz}}(n)}{\phi_{y_{iz}}^{\mathrm{MVDR}}(n) - \phi_{s_{iz}}(n)} \\ &= \frac{\mathrm{SIX}_{iz}^{\mathrm{PCMV}}(n)}{1 - \mathrm{SIX}_{iz}^{\mathrm{PCMV}}(n)}, \end{aligned} \tag{17}$$

[2] The mixing rule can be chosen arbitrarily as long as the resulting mixing vectors are linearly independent. The above mixing rule is simple and also allows for a geometric interpretation.

[3] Algorithm 1 uses the MVDR filter from (4) in (6) to assess ϕ_{y_i}.

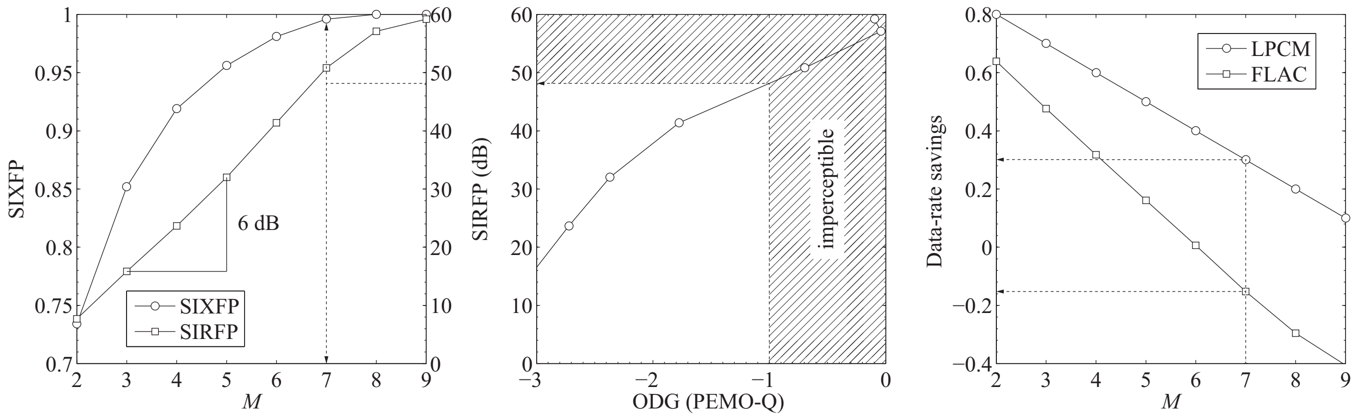

Fig. 2. Mean SIXFP, SIRFP and ODG for the multitrack and the corresponding data-rate savings. Starting from the middle, we see that for imperceptible quality impairment, i.e. an ODG value above -1, the SIRFP value must be 48 or greater (shaded area). Switching over to the left, we see that at least 7 channels are necessary to reach it. The figure on the right indicates that 30 % of LPCM data can so be saved.

which is invertible. For numerical reasons, however, it is advisable to convert SIX to $\mathrm{SIR}^{\mathrm{out}}$ first, and to limit the range of $\mathrm{SIR}^{\mathrm{out}}$ to ± 60 dB afterwards. By weighting the SIX metric by frequency and fractional input power we obtain another metric:

$$\mathrm{SIXFP}_i(n) = \frac{\sum_z \phi_{s_{iz}}(n)\mathrm{SIX}_{iz}(n)}{\sum_z \phi_{s_{iz}}(n)}. \qquad (18)$$

In the case of the PCMV filter, $\mathrm{SIRFP}_i^{\mathrm{out}}(n)$ can also be computed from $\mathrm{SIXFP}_i(n)$ using (17). The overall average is calculated as the arithmetic mean over the time segments in which the composite input signal power is significant:

$$\mathrm{SIXFP}_i = \frac{1}{|\mathsf{N}|}\sum_{n\in\mathsf{N}} \mathrm{SIXFP}_i(n), \qquad (19)$$

where $\mathsf{N} = \left\{n \mid \sum_z \phi_{s_{iz}}(n) \geqslant \phi\right\}$ and ϕ is an empirically chosen lower bound.

4.2. Control mechanism

As SIXFP is a measure of similarity between the original and the estimated components, it can be used to predict the signal quality at the output before the final mixdown. For this, the local covariance matrix in (15) is computed as in (3) from the quantized STPSDs that are available after analysis and the tentative mixing coefficients. One starts with the lowest possible value for M, which is 2, and increases M until the desired SIXFP value is reached. For $M = I$, perfect reconstruction is expected. The stop condition can be defined globally for the entire signal or locally for a segment. One can also have a single condition for all objects or a separate condition for each one of them.

5. PERFORMANCE EVALUATION

5.1. Experimental design

We use the testing framework from [8]. The number of frequency bands is set to 76, which results in a mean side-information rate of 11.5 kbps per object at a sampling rate of 44.1 kHz. The proposed scheme is tested on Fort Minor's "Remember the Name" 10-track recording of 20 s length. All tracks are converted to mono. FLAC is used to code the mixture, which is not watermarked. The resulting audio quality is evaluated for 2–9 mixture channels.

5.2. Experimental results

The results are shown in Fig. 2. The accompanying sound clips can be downloaded from http://www.labri.fr/˜gorlow/icassp13/. It can be seen that for imperceptible quality impairment correspondent to PEMO-Q's ODG metric [14, 15], one requires that $\mathrm{SIXFP} \geqslant 0.99$ or $\mathrm{SIRFP} \geqslant 48.0$ dB. This corresponds to 7 channels for the given multitrack. It can further be noted that $\Delta\mathrm{SIRFP} \approx 6 \cdot \Delta M$, i.e. the SIRFP value increases approximately by 6 dB with each additional channel. The data-rate savings due to *downmixing* equal $1 - M/I$. They amount to 0.3 in the above example, see the LPCM curve. Yet the lower curve conveys that coding the 10 mono tracks with FLAC *separately* is more efficient than coding the 7 channels, i.e. so long as *interchannel redundancy* is not minimized. Even so, according to informal listening tests, perceptual transparency is already attained with 5 channels. In that case, the proposed scheme provides savings of 0.5 for the uncoded LPCM mixture, or 0.2 when it is coded with FLAC. The ratio of side information to FLAC-coded data is 0.14 or less, and scales with the channel number M.

6. CONCLUSION

In this work, we introduced a lossy coding scheme that can reduce the storage capacity for a multitrack recording. In virtue of a quality control mechanism, which is inherent in the proposed scheme, it is possible to scale the resulting audio quality as a function of data-rate savings and vice versa. For the exemplary multitrack, we were able to reduce the raw data by 30 %, attaining perceptual transparency. It should be possible to achieve a higher percentage if the redundancy between channels is also taken into account.

In addition, it can be noted that the computation of the proposed quality metrics is by far less costly than the utilization of algorithms that simulate perceptual properties of the human ear, such as PEAQ [16] or PEMO-Q. A direction for future work is thus to find a direct mapping between any of the two metrics and, e.g., the ODG. It may also be worth mentioning that spatial filtering is likewise applicable to any legacy multichannel object-based format which incorporates linearly independent mixing vectors.

7. REFERENCES

[1] C. Faller, "Parametric joint-coding of audio sources," in *AES Convention 120*, May 2006.

[2] J. Engdegård *et al.*, "Spatial Audio Object Coding (SAOC) — the upcoming MPEG standard on parametric object based audio coding," in *AES Convention 124*, May 2008.

[3] ISO/IEC, *Information technology — MPEG audio technologies — Part 2: Spatial Audio Object Coding (SAOC)*, Oct. 2010, ISO/IEC 23003-2:2010.

[4] J. Herre *et al.*, "MPEG Spatial Audio Object Coding — the ISO/MPEG standard for efficient coding of interactive audio scenes," *J. Audio Eng. Soc.*, vol. 60, no. 9, pp. 655–673, Sep. 2012.

[5] G. Hotho, L. F. Villemoes, and J. Breebaart, "A backward-compatible multichannel audio codec," *IEEE Trans. Audio Speech Lang. Process.*, vol. 16, no. 1, pp. 83–93, Jan. 2008.

[6] K. H. Knuth, "Informed source separation: A Bayesian tutorial," in *Proc. EUSIPCO*, 2005.

[7] S. Gorlow and S. Marchand, "Informed source separation: Underdetermined source signal recovery from an instantaneous stereo mixture," in *Proc. IEEE WASPAA*, 2011, pp. 309–312.

[8] ——, "Informed audio source separation using linearly constrained spatial filters," *IEEE Trans. Audio Speech Lang. Process.*, vol. 21, no. 1, pp. 3–13, Jan. 2013.

[9] "Free Lossless Audio Codec (FLAC)," http://flac.sourceforge.net, version 1.2.1b.

[10] R. Geiger, Y. Yokotani, and G. Schuller, "Audio data hiding with high rates based on IntMDCT," in *Proc. IEEE ICASSP*, 2006, pp. 205–208.

[11] J. Pinel and L. Girin, "A high-rate data hiding technique for audio signals based on IntMDCT quantization," in *Proc. DAFx*, 2011, pp. 353–356.

[12] H. Cox, R. M. Zeskind, and M. M. Owen, "Robust adaptive beamforming," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 35, no. 10, pp. 1365–1376, Oct. 1987.

[13] J. Capon, "High-resolution frequency-wavenumber spectrum analysis," in *Proc. IEEE*, 1969, pp. 1408–1418.

[14] R. Huber and B. Kollmeier, "PEMO-Q — a new method for objective audio quality assessment using a model of auditory perception," *IEEE Trans. Audio Speech Lang. Process.*, vol. 14, no. 6, pp. 1902–1911, Nov. 2006.

[15] HörTech gGmbH, "PEMO-Q," http://www.hoertech.de/web_en/produkte/pemo-q.shtml, version 1.3.

[16] ITU-R, *Method for objective measurements of perceived audio quality*, Nov. 2001, rec. ITU-R BS.1387-1.