On the Informed Source Separation Approach for Interactive Remixing in Stereo

Stanislaw Gorlow, Sylvain Marchand

Audio Engineering Society
Convention Paper
Presented at the 134th Convention
2013 May 4–7 Rome, Italy

On the Informed Source Separation Approach for Interactive Remixing in Stereo*

Stanislaw Gorlow[1] and Sylvain Marchand[2]

[1]*Univ. Bordeaux, LaBRI, UMR 5800, 33400 Talence, France*

[2]*Univ. Brest, Lab-STICC — CNRS UMR 6285, 29238 Brest, France*

Correspondence should be addressed to Stanislaw Gorlow (stanislaw.gorlow@labri.fr)

ABSTRACT

Informed source separation (ISS) has become a popular trend in the audio signal processing community over the past few years. Its purpose is to decompose a mixture signal into its constituent parts at the desired or the best possible quality level given some metadata. In this paper we present a comparison between two ISS systems and relate the ISS approach in various configurations with conventional coding of separate tracks for interactive remixing in stereo. The compared systems are Underdetermined Source Signal Recovery (USSR) and Enhanced Audio Object Separation (EAOS). The latter forms a part of MPEG's Spatial Audio Object Coding technology. The performance is evaluated using objective difference grades computed with PEMO-Q. The results suggest that USSR performs perceptually better than EOAS and has a lower computational complexity.

1. BACKGROUND

About a decade ago, in 2003, Avendano presented a scheme [1] similar to [2] with which one could identify, separate, and manipulate individual sound sources in a studio music recording. His technique uses a "panning index" to identify collinear source signal components and clusters those into coherent time-frequency regions [3, 4]. These regions can then be manipulated by applying a "mask" that alters the magnitude of the signal components in question. In that manner one can either attenuate or accentuate the source of interest—the vocal or an instrument—and even change its apparent location. These features, which are known as *karaoke*, *mix-minus*, or *repanning* otherwise, are all basic elements of *active listening* [5, 6]. Avendano's technique, which is applicable to convolutional stereo mixtures without any restrictions with regard to the number of superposed sound sources, has one drawback: the resulting audio quality is insufficient for professional audio applications. A simi-

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

lar technique for *binaural* recordings was developed by Mouba and Marchand in 2006 [7].

In order to attain a higher quality as compared to Avendano, Oh *et al.* presented in [8] a *model-based* remixing scheme that likewise allows for gain manipulations and repanning of distinct sound sources, but with the aid of some *additional information* including the model/mixing parameters and the approximate short-time power spectral densities (STPSDs) of the sound sources that are to be manipulated. The additional information, which is transmitted alongside a stereo mixture signal, is used to best fit the re-/mixing model in the least-squares (LS) sense given new user-definable mixing parameters. The authors claim their technique to require less side information than other comparable schemes such as MPEG Spatial Audio Object Coding (SAOC) [9–11] to achieve the same effect, as only the STPSDs of the few selected sound sources and their mixing coefficients need to be communicated to the remixing unit. However, if the user was intended to be given the possibility to alter the entire mix, the required amount of side information would coincide with the one of SAOC.

In [12], Knuth gives a Bayesian tutorial on the design of robust algorithms for source separation that take advantage of *prior information* about the problem in order to assure that one reaches an optimal solution. Many of today's state-of-the-art score-informed source separation techniques are Bayesian inference based, see e.g. [13–15]. Most commonly, they make use of one of many variants of the nonnegative matrix factorization (NMF) [16] to approximate the spectrograms of the original signals, so as to apply masks in the form of linear filters to the mixture spectrum. The source signal components are hence estimated as the means of the respective *posterior distributions* using the mean-square error (MSE) as risk. The linear filters in this case are minimum-mean-square-error (MMSE) estimators.

ISS can be seen as a new formalism for a known coding paradigm [17–19], for it has the following advantages. It has a modular framework which is downward compatible with mono and stereo on the one hand, but also upward extensible to multichannel on the other hand. The linear MMSE estimator, for example, can be generalized to an arbitrary number of mixture channels. Furthermore, ISS is "future improvable", since it supports different kinds of estimators so long as those follow one and the same paradigm.

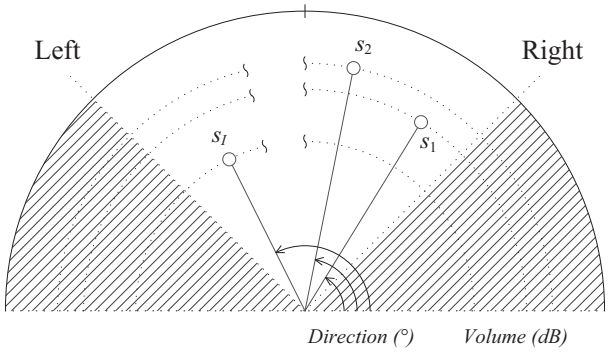

Fig. 1: Modeling of sound sources in a sound field using the parameters *direction* (azimuth) and *volume* (radius).

The aim of this paper is to illustrate the ISS approach, to identify similarities between our ISS system and SAOC, and to provide some objective performance data of our system in a remixing scenario given a corpus of realistic multitrack recordings. We also want to relate the results to what is achievable with conventional coding schemes. The remaining part of the paper is organized as follows. Section 2 presents the mixture model upon which our considerations are based. The ISS approach is explained in Section 3 by giving two examples of state-of-the art ISS systems: ours and an SAOC subsystem. We compare the two systems and investigate the performance of our system in a remixing scenario in Section 4. Conclusions are drawn in Section 5.

2. STEREOPHONIC MIXTURE MODEL

Consider a stereo mixing system in which one or multiple mono signals are unevenly distributed over two independent audio channels in such a way as to create an illusion of directionality and audible perspective. This is achieved by varying the amplitude of the signal sent to each channel relative to the listener. The parameters that control this relative amplitude are *direction* and *volume* (see Fig. 1). They are equivalent to the position of the panoramic potentiometer, the pan-pot, and the fader position on a mixing desk in a recording studio and are applied to each mono signal separately. The summation of all pan-potted and volume adjusted mono signals constitutes the final sound field of the mixture. Accordingly, our data model in the discrete Fourier transform (DFT)

domain is

$$\mathbf{X}^{\mathsf{H}} = \sum_{i=1}^{I} \mathbf{a}_i \mathbf{s}_i^{\mathsf{H}} = \mathbf{A}\mathbf{S}^{\mathsf{H}}, \tag{1}$$

where $\mathbf{s}_i \in \mathbb{C}^N$ is a N-length signal vector, $\mathbf{a}_i \in \mathbb{R}^2$ is the mixing vector that represents the spread of the ith source signal into the sound field, I is the total number of sound sources, and $\mathbf{X} \in \mathbb{C}^{N\times 2}$ is the mixture signal. Superscript H denotes Hermitian transpose. The I signal vectors and the I mixing vectors can be concatenated into the signal matrix $\mathbf{S} \in \mathbb{C}^{N\times I}$ and the mixing matrix $\mathbf{A} \in \mathbb{R}^{2\times I}$, such that the mixture matrix can be expressed as a product of the two matrices. Since I is always larger than two, the mixing process may also be termed as "downmix" or "fold-down". In respect of the mixing vector, we want to restrict ourselves to the case where each source signal was already preadjusted to the desired volume level $b_i \in \mathbb{R}$ as per

$$\mathbf{s}_i = b_i \mathbf{s}_i', \tag{2}$$

where $\mathbf{s}_i' \in \mathbb{C}^N$ is the recorded signal normalized to a nominal level of say -16 dBFS. The mixing vector $\mathbf{a}_i$ dictates how the signal power is distributed between left and right and is defined as

$$\mathbf{a}_i \triangleq \begin{bmatrix} \sin\theta_i & \cos\theta_i \end{bmatrix}^{\mathsf{T}}, \tag{3}$$

so that $\|\mathbf{a}_i\| = 1\,\forall\theta_i \in [0°, 90°]$. This implies that the signal power is kept constant regardless of the angle. The angle range is defined in such a manner that at the lower end, 0°, the source appears in only the right channel. On the other hand, when placed at the upper end, 90°, the source appears in only the left channel. In the middle, 45°, the signal power is equally distributed across the two output channels and the source appears in a phantom center channel.

3. INFORMED AUDIO SOURCE SEPARATION

3.1. Outline

The primary objective of informed source separation is to separate the components of a given mixture. This can be accomplished with statistical methods like Bayesian inference or optimum filtering, which are two equivalent approaches when using the MSE as risk. These methods are of particular interest because they assume previous knowledge of the source parameters and their statistical behavior, which can be described in terms of the mean and variance. They also require an explicit model of the mixture including the mixing parameters to estimate the original signals. Imprecise or limited knowledge of any of the parameters has direct impact on the performance of the estimator. The second objective is thus to provide some additional information for the estimator so that its accuracy is improved, which leaves a margin for quality control. The separator does not necessarily have to be a Bayes estimator or an optimum filter. It can also include any pre- or post-processing that would boost the quality beyond the limits of a standalone estimator.

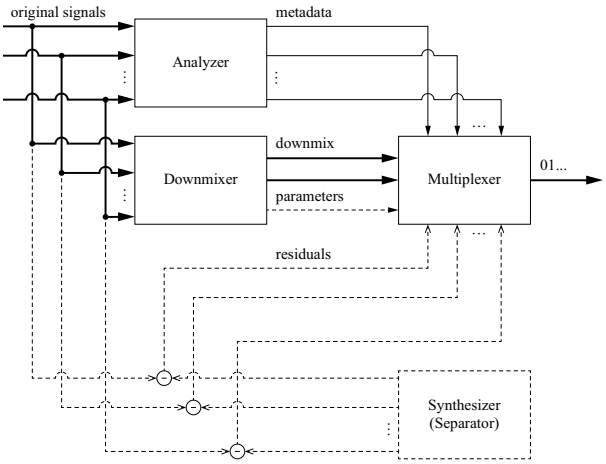

Fig. 2: Generic ISS encoder.

In a concrete realization of an ISS system, in which the metadata is extracted from the source signals but those are no longer accessible after they have been mixed, it is mandatory to uncouple the extraction phase from the separation phase. The content creator must provide via an encoder all the required information for the content consumer to decompose the mixture via a decoder. The task of the encoder is to extract a minimum of ancillary data from the source signals, so the decoder can recover copies from the mixture in high perceptual quality. The metadata can either be hidden in the mixture signal in the form of an inaudible watermark [20, 21] or must be embedded as additional data within the bitstream.

Fig. 2 shows a generic scheme for a practical encoder. It contains an *analyzer* block that extracts the metadata from the original signals, a *downmixer* block that represents the mixing system, a *multiplexer* block that assembles the bitstream, and an optional and hence dashed *synthesizer* block. Its presence is justified by the fact that the separator commonly exhibits an upper limit of

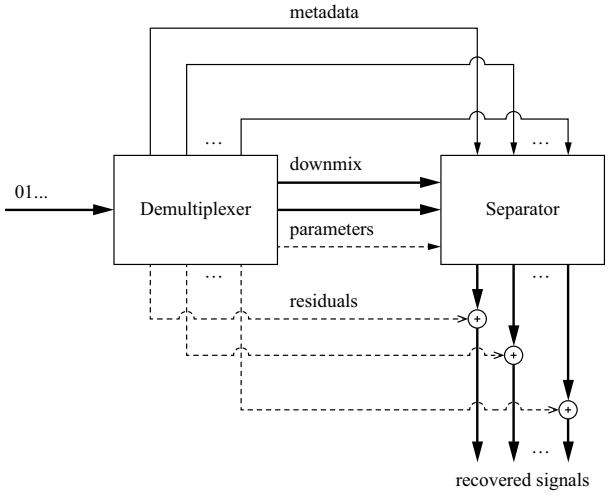

Fig. 3: Generic ISS decoder.

performance. A straightforward way to improve the estimates is to code the corresponding residuals using an analysis-by-synthesis approach. As has been demonstrated in [22], the code rate of the residuals can be adjusted in such a way as to guarantee a desired output quality level.

The associated decoder is shown in Fig. 3. A *demultiplexer* disassembles the bitstream into the downmix and the metadata plus the residuals in case these are available. The separator then estimates the original signals from the downmix using the metadata. The estimated signals, which are optionally corrected by the residuals, represent the recovered signals.

3.2. Underdetermined Source Signal Recovery

3.2.1. Brief overview

Underdetermined Source Signal Recovery (USSR) [23, 24] is an example of an ISS system. Others can be found in [25–27] and there exist surely more. A comparison of the cited systems, or rather their derivatives, has shown that USSR performs perceptually best for both an instantaneous and a convolutional stereo mixture [28].

USSR is as a subband-domain spatial filtering technique that uses short-time power spectral densities (STPSDs) of the original source signals as metadata. The STPSDs serve to constrain the output signal power and to model the spatial correlation between the sources in each point of the time-frequency plane. To reduce the metadata rate, the STPSDs are quantized on a log-log scale (see 3.2.2).

With USSR, the superposed source signals are spatially decorrelated and their estimates are constrained to have the desired power level to overcome the issue of spectral gaps occurring in some corpus of other techniques.

3.2.2. Encoder

As the USSR system operates in the short-time Fourier transform (STFT) domain, the source signal matrix $\mathbf{S}$ in (1) holds the complex Fourier coefficients of the source signals for the duration of a time segment in an arbitrary frequency band. The band size N relates to an ERB-like frequency scale given by

$$\mathrm{ERBS}(f) = \lfloor 21.4 \log_{10}(4.37f + 1) \rfloor, \tag{4}$$

where f is in kHz and $\lfloor \cdot \rfloor$ represents the floor function. The STPSDs are averaged over the "critical" bands and quantized according to

$$\tilde{\phi}_{i,\log} = \left\lfloor 5 \log_{10} \tfrac{1}{N} \mathbf{s}_i^{\mathsf{H}} \mathbf{s}_i \right\rceil, \tag{5}$$

where $\lfloor \cdot \rceil$ represents the rounding function. These power values are differentially pulse-code modulated in either the time or frequency direction and entropy coded.

3.2.3. Decoder

The core of the decoder and the overall system is formed by the so-called "power-conserving minimum-variance" (PCMV) filter. Its weights are calculated as

$$\hat{\mathbf{w}}_i^{\mathrm{PCMV}} = \sqrt{\frac{\tilde{\phi}_i}{\mathbf{a}_i^{\mathsf{T}} \hat{\mathbf{R}}_{\mathbf{xx}}^{-1} \mathbf{a}_i}} \hat{\mathbf{R}}_{\mathbf{xx}}^{-1} \mathbf{a}_i, \tag{6}$$

where $\hat{\mathbf{R}}_{\mathbf{xx}}$ is an estimate for the spatial correlation matrix of the observed mixture, which is approximated as

$$\hat{\mathbf{R}}_{\mathbf{xx}} \approx \mathbf{A} \mathrm{diag}\left(\tilde{\phi}_1, \tilde{\phi}_2, \ldots, \tilde{\phi}_I\right) \mathbf{A}^{\mathsf{T}}. \tag{7}$$

Equation (7) states that the source signals are mutually independent and thus uncorrelated—entries outside the main diagonal are all zero. The mixing vectors $\{\mathbf{a}_i\}$ are deduced from the corresponding angles $\{\theta_i\}$, which are either estimated from the mixture or transmitted. Lastly, the source signals are estimated as

$$\hat{\mathbf{s}}_i = \mathbf{X} \hat{\mathbf{w}}_i^{\mathrm{PCMV}}. \tag{8}$$

3.3. Enhanced Audio Object Separation

3.3.1. Brief overview

SAOC's own source separation scheme is referred to as "enhanced audio object separation" (EAOS), whereas the

object decoder is also called an "enhanced audio object processor" [29, Fig. 6]. The latter consists of a two-to-I upmix unit [29, Fig. 5] which corresponds to I two-to-three (TTT) basic units [29, Fig. 4]. A TTT unit takes a two-channel mixture as input and outputs the mono signal from the object of interest, the *foreground object* (FGO), in conjunction with the remaining background signal which is the sum of all pan-potted object signals excluding the FGO signal. Every object that contributes to the background signal is further deemed a *background object* (BGO). A BGO is unalterable, while an FGO can be altered in regard to its volume level and location. In applications like karaoke or mix-minus, where a single track is soloed or muted, an FGO is upgraded to a so-called "enhanced audio object" (EAO). An EAO signal is tantamount to an FGO signal that is error corrected and as such quality improved using the coded residual at the cost of a higher side-information rate.

3.3.2. Encoder

Like USSR, EAOS too analyzes the source signals in the complex subband domain. The subband analysis is based on a hybrid filter bank splitting the time signal into 69 subband signals [30]. The STPSDs are obtained from the instantaneous powers in each subband, which are quantized on a logarithmic scale and grouped over time and frequency. Finally, this metadata is differential-entropy coded and passed on to the decoder as side information along with the mixing coefficients (b_i, θ_i) and the coded downmix signal. The calculation of a single residual is as follows. First, all BGO signals are combined into a downmix signal [9, Eqs. 15–16]

$$\mathbf{X}_{\mathrm{BGO}}^{\mathsf{H}} \triangleq \mathbf{X}^{\mathsf{H}} - \mathbf{a}_{\mathrm{FGO}}\mathbf{s}_{\mathrm{FGO}}^{\mathsf{H}}. \tag{9}$$

Second, an "auxiliary" signal is calculated by taking the difference between the FGO and the BGO mixture signal that was projected onto the "look direction" of the FGO according to [9, Eqs. 15–16]

$$\mathbf{s}_{\mathrm{FGO}_\circ} \triangleq \mathbf{X}_{\mathrm{BGO}}\mathbf{a}_{\mathrm{FGO}} - \mathbf{s}_{\mathrm{FGO}}. \tag{10}$$

Using (9) and (3), (10) can also be written as

$$\mathbf{s}_{\mathrm{FGO}_\circ} = \mathbf{X}\mathbf{a}_{\mathrm{FGO}} - 2\mathbf{s}_{\mathrm{FGO}}. \tag{11}$$

Then, a linear combination of the two downmix channels is found that minimizes the reconstruction error between the modeled signal and the true $\mathrm{FGO}_\circ$ signal. For this, a system of N linear equations in two unknowns $\mathbf{w}_{\mathrm{FGO}_\circ} = \left[w_{\mathrm{FGO}_\circ,1}\ w_{\mathrm{FGO}_\circ,2}\right]^{\mathsf{T}} \in \mathbb{R}^2$ must be solved:

$$\mathbf{X}\mathbf{w}_{\mathrm{FGO}_\circ} = \mathbf{s}_{\mathrm{FGO}_\circ}. \tag{12}$$

The weight coefficients that best fit the above equations in the LS sense are

$$\hat{\mathbf{w}}_{\mathrm{FGO}_\circ}^{\mathrm{LS}} = \underbrace{\left(\mathbf{X}^{\mathsf{H}}\mathbf{X}\right)}_{\hat{\mathbf{R}}_{\mathbf{xx}}}^{-1}\underbrace{\mathbf{X}^{\mathsf{H}}\mathbf{s}_{\mathrm{FGO}_\circ}}_{\hat{\mathbf{p}}_{\mathbf{x}\mathbf{s}_{\mathrm{FGO}_\circ}}}. \tag{13}$$

The terms $\mathbf{X}^{\mathsf{H}}\mathbf{X}$ and $\mathbf{X}^{\mathsf{H}}\mathbf{s}_{\mathrm{FGO}_\circ}$ in (13) are tantamount to the sample estimates of spatial correlation, $\hat{\mathbf{R}}_{\mathbf{xx}}$ and $\hat{\mathbf{p}}_{\mathbf{x}\mathbf{s}_{\mathrm{FGO}_\circ}}$. For that reason, the weighted LS estimate for $\mathbf{w}_{\mathrm{FGO}_\circ}$ is formally identical with the MMSE estimator for the auxiliary signal component $\mathrm{FGO}_\circ$. Due to (11),

$$\hat{\mathbf{p}}_{\mathbf{x}\mathbf{s}_{\mathrm{FGO}_\circ}} = \hat{\mathbf{R}}_{\mathbf{xx}}\mathbf{a}_{\mathrm{FGO}} - 2\hat{\mathbf{p}}_{\mathbf{x}\mathbf{s}_{\mathrm{FGO}}}, \tag{14}$$

so by rewriting (13) as

$$\hat{\mathbf{w}}_{\mathrm{FGO}_\circ}^{\mathrm{MMSE}} = \mathbf{a}_{\mathrm{FGO}} - 2\underbrace{\hat{\mathbf{R}}_{\mathbf{xx}}^{-1}\hat{\mathbf{p}}_{\mathbf{x}\mathbf{s}_{\mathrm{FGO}}}}_{\hat{\mathbf{w}}_{\mathrm{FGO}}^{\mathrm{MMSE}}}, \tag{15}$$

the estimator can be put in direct relation to the FGO, where $\mathbf{p}_{\mathbf{x}\mathbf{s}_{\mathrm{FGO}}}$ represents the cross-correlation between the mixture and the FGO signal. The difference between the true $\mathrm{FGO}_\circ$ signal and its estimate,

$$\mathbf{r}_{\mathrm{FGO}_\circ} = \mathbf{s}_{\mathrm{FGO}_\circ} - \mathbf{X}\hat{\mathbf{w}}_{\mathrm{FGO}_\circ}^{\mathrm{MMSE}}, \tag{16}$$

gives the residual that is perceptual-entropy coded using the Advanced Audio Coding (AAC) [31] scheme at an average bitrate of 20 kbps [29].

3.3.3. Decoder

In order to obtain the FGO signal, one needs to estimate the auxiliary signal first. To that end, one calculates the estimator coefficients in (15) using the power spectra and the mixing coefficients. Both have been made available to the decoder by the encoder. The correlation matrix is calculated as in (7), whereas the cross-correlation vector is approximated as

$$\hat{\mathbf{p}}_{\mathbf{x}\mathbf{s}_{\mathrm{FGO}}} \approx \mathbf{a}_{\mathrm{FGO}}\tilde{\phi}_{\mathrm{FGO}}. \tag{17}$$

Here again, the audio objects are deemed to be mutually uncorrelated. Note that in SAOC the cross-correlations between audio objects can also be approximated using the "inter-object coherences". The transmission of these is optional. Furthermore, according to our experience, nonzero cross-correlation terms decrease the quality of

estimates rather than improving it. Once the estimator has been calculated, it is plugged into (12), so that the enhanced $\mathrm{FGO}_\circ$ signal is obtained by adding the residual to the estimate:

$$\tilde{\mathbf{s}}_{\mathrm{EAO}_\circ} = \underbrace{\mathbf{X}\hat{\mathbf{w}}_{\mathrm{FGO}_\circ}^{\mathrm{MMSE}}}_{\hat{\mathbf{s}}_{\mathrm{FGO}_\circ}} + \tilde{\mathbf{r}}_{\mathrm{FGO}_\circ}. \tag{18}$$

Solving (10) for $\mathbf{s}_{\mathrm{FGO}}$ using (9) yields the sought-after EAO signal

$$\tilde{\mathbf{s}}_{\mathrm{EAO}} = \tfrac{1}{2}(\mathbf{X}\mathbf{a}_{\mathrm{FGO}} - \tilde{\mathbf{s}}_{\mathrm{EAO}_\circ}). \tag{19}$$

Using (18) and (15), (19) can also be formulated as

$$\tilde{\mathbf{s}}_{\mathrm{EAO}} = \mathbf{X}\hat{\mathbf{w}}_{\mathrm{FGO}}^{\mathrm{MMSE}} - \tfrac{1}{2}\tilde{\mathbf{r}}_{\mathrm{FGO}_\circ}. \tag{20}$$

The bottom line of (20) is that the TTT unit in SAOC is a weighted LS/MMSE estimator for the FGO signal with a particular residual coding strategy. Such being the case, it fits perfectly into the ISS framework from Figs. 2–3.

4. PERFORMANCE EVALUATION

4.1. Underdetermined Source Signal Recovery Versus Enhanced Audio Object Separation

In the previous section it has been shown that EAOS in SAOC uses a Bayes estimator in the form of an MMSE spatial filter to separate audio objects given the mixture. In this section we compare EAOS with USSR using the same testing framework.

4.1.1. Test setup

For both systems we use the STFT with a Kaiser-Bessel derived window and 50-% overlap between contiguous time segments. The PCMV estimator in USSR is replaced by the TTT unit when simulating EAOS. The metadata is encoded as in [24]. The mixing coefficients are considered to be known. The two systems are compared with each other in terms of quality and computational complexity. The quality is assessed on the "objective difference grade" (ODG) scale [32], while the complexity is measured by the execution time of the decoder in MATLAB. The ODG score is computed with the PEMO-Q software [33, 34]. Ten multitracks are taken from the QUASI database [35], converted to mono and cut down to 20-s excerpts. Each track is normalized to a reference root-mean-square (RMS) level of -16 dBFS. The audio sources are placed uniformly in space and gain

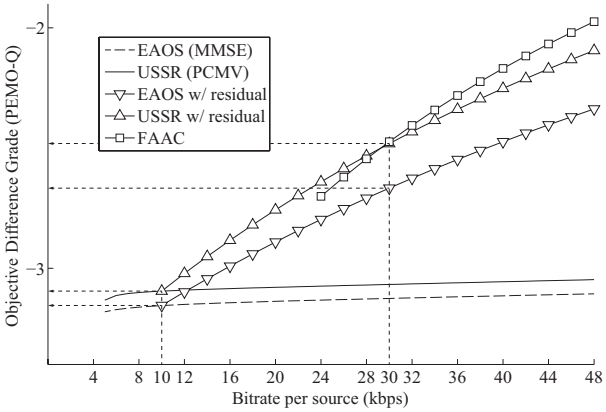

Fig. 4: ODG as a function of the bitrate.

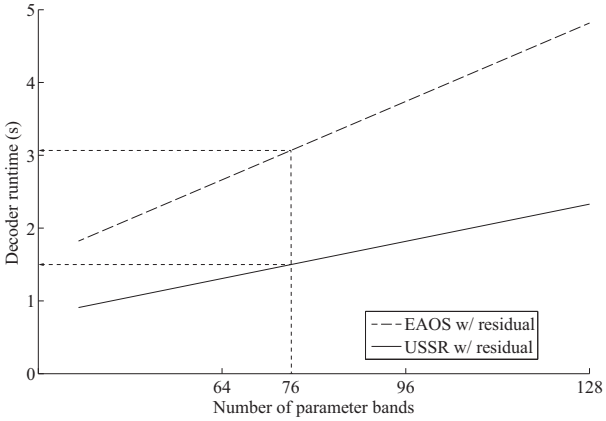

Fig. 5: Decoder runtime as a function of the number of bands at 30-kbps side-information rate.

adjusted, so as to have an equal signal-to-interference ratio across all sources at the output.

4.1.2. Test results

The results of the comparison are shown in Figs. 4–5. In the case where the residual is omitted, the bitrate is equivalent to the metadata rate for a varied number of parameter bands. In the case where the residual is used to correct the initial estimate, the metadata rate is fixed at 10 kbps and the bitrate is calculated as the sum of the latter and the residual rate which is increased from zero onwards. We also include the results where the original signals are coded separately. There, the bitrate is simply the coding rate.

In our experiment we use the Freeware Advanced Audio Coder (FAAC) [36] for the residual and the original tracks. The results are averaged over the complete data corpus. As can be seen from Fig. 4, USSR's PCMV estimator yields better results than EAOS's LS estimator. This observation is consistent with the listening test results reported in [24] in which the PCMV estimator outperforms the MMSE estimator. The gap between the two systems widens even further if their estimates are error corrected. At a bitrate of roughly 30 kbps per source and beyond, the quality of FAAC-coded tracks is superior to that of the tracks recovered from the mixture.

As shown in Fig. 5, the USSR decoder is approximately half as complex as the EAOS decoder if all I estimates are error corrected. The longer runtime is explained by the fact that EAOS requires the residuals to be available in the subband domain, whereas USSR does not. Hence, the loss of time is due to the extra I-fold STFT and the computation of auxiliary signals.

4.2. Interactive Remixing

In this section we evaluate the informed approach in a concrete scenario. We choose "interactive remixing" as sample application which allows the user to change the volume level of sound sources and their spatial location. We are also interested in finding out whether "plain" source coding is a more pragmatic solution in terms of audio quality and coding efficiency.

4.2.1. Test setup

We use the same testing framework as before. The ODG is retained as quality index. In order to simulate more realistic conditions, the stereo-to-mono converted tracks are panned to their original location, so the downmix is prearranged as if by the sound engineer or the composer. The location is derived from the RMS channel difference in the original stereo tracks:

$$\hat{\theta}_i = \operatorname{arccot} \frac{\mathrm{RMS}_{i,\mathrm{right}}}{\mathrm{RMS}_{i,\mathrm{left}}}, \tag{21}$$

where arccot is the arccotangent.

Table 1 provides a listing of the songs used in the experiment. The source signals are either recovered from the downmix using USSR or are encoded (and decoded) for comparison. For this, we use FAAC and the Enhanced aacPlus [37, 38] coder. The downmix is encoded in perceptually transparent quality at a variable bitrate of approximately 120 kbps. To simulate user interaction, we generate ten different remixes with arbitrary new source locations and volume levels for each song and system with the gains being in the range between -6 and 3 dB. We then compare the remixes of each system with the ones created from the original tracks.

It should be noted that the evaluation software PEMO-Q does not take spatial hearing effects into account. Hence, the *subjective* quality can be expected to be higher. The assessed *relative order* of systems should yet remain the same in either case.

4.2.2. Test results

The evaluation results are summarized in Fig. 6. It can be observed that the quality of a remix that is created from an FAAC-coded downmix depends on the number of sources in the mix as much as their spatial spread. A linearly pulse-code modulated (LPCM) mixture signal seems less sensitive to these factors. The deciding factor there is apparently the spectral texture of a source signal and by how much it interweaves with other sources.

On average, the best quality is obtained for an LPCM mixture in combination with residual coding at roughly 20 kbps per sound source. With a median not worse than "slightly annoying", the results gained with USSR alone are also promising. Clearly worse are the results for the scenario in which the mixture is FAAC coded. Even at a side-information rate of 30 kbps, the quality lies halfway between "annoying" and "slightly annoying". The same is true for Enhanced aacPlus at 10 kbps or FAAC at 30 kbps. The most efficient system in the experiment is Enhanced aacPlus operating at 30 kbps, because it does not necessitate availability of the mixture.

Song no.	Title	Number of sources	Spatial spread vs. centroid
1	“Carol of the Bells” (Alexq)	4	15.9° / 40.3°
2	“One We Love” (Another Dreamer)	5	20.8° / 47.4°
3	“The World Is Under Attack” (Carl Leth)	6	5.76° / 47.2°
4	“Remember the Name” (Fort Minor)	10	10.4° / 46.4°
5	“The Spirit of Shackleton” (Glen Philips)	12	7.54° / 47.1°
6	“Mix Tape” (Jim’s Big Ego)	7	2.00° / 45.0°
7	“Good Soldier” (Nine Inch Nails)	5	1.72° / 43.8°
8	“Sunrise” (Shannon Hurley)	8	8.51° / 41.0°
9	“Untitled” (Ultimate NZ Tour)	7	12.3° / 45.3°
10	“Ana” (Vieux Farka)	8	9.54° / 42.6°

Table 1: The corpus of prearranged mixes.

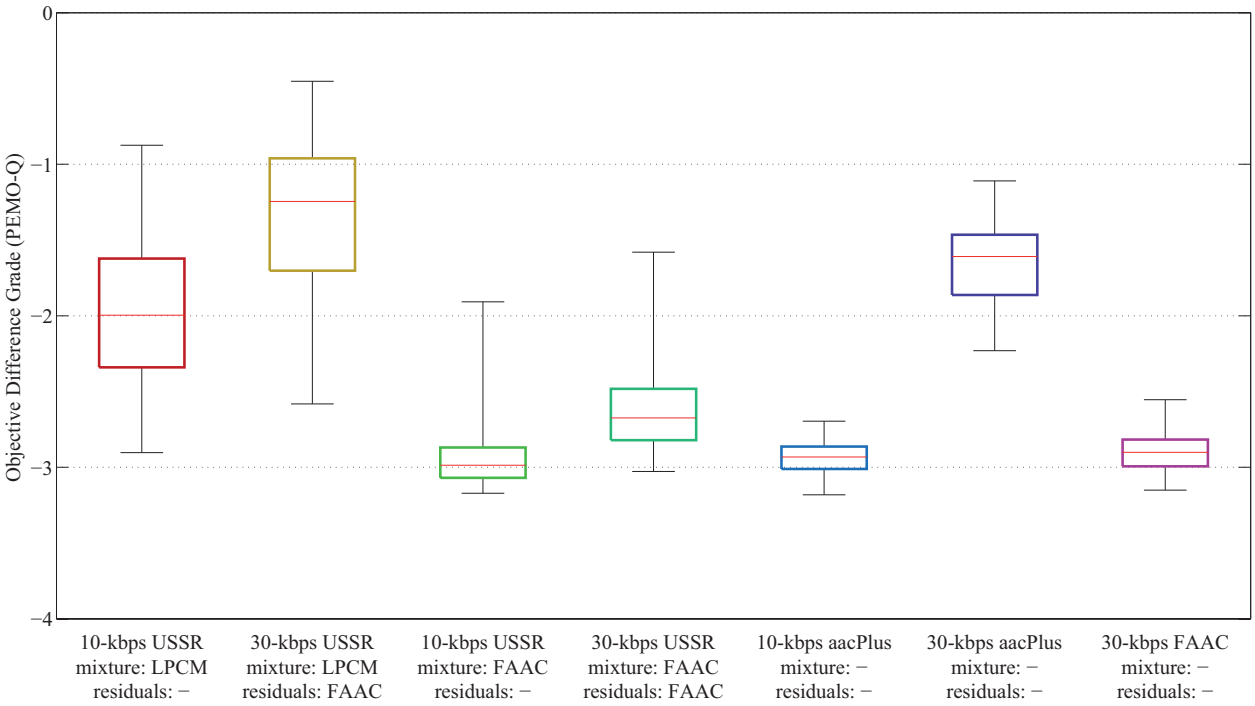

Fig. 6: The medians and the 25th and 75th percentiles for each system under test.

5. CONCLUSION

We attempted to give the reader a better understanding of the informed source separation approach by contrasting it with MPEG's technology for interactive remixing: SAOC. We demonstrated that the enhanced audio object processor is an MMSE estimator with error correction that perfectly fits into the ISS framework. The fact that USSR and EAOS both operate in the subband domain and also use the same metadata, allowed us to use the two techniques in a single implementation. In this vein we were able to compare their performance. The results indicate that USSR performs better than EAOS. Also, it was observed that the Enhanced aacPlus coder provides a better remix quality in comparison to when USSR is applied to an FAAC-coded mixture.

6. REFERENCES

[1] C. Avendano, "Frequency-domain source identification and manipulation in stereo mixes for enhancement, suppression and re-panning applications," in *Proc. IEEE WASPAA*, 2003, pp. 55–58.

[2] A. Jourjine, S. Rickard, and O. Yılmaz, "Blind separation of disjoint orthogonal signals: Demixing N sources from 2 mixtures," in *Proc. IEEE ICASSP*, vol. 5, 2000, pp. 2985–2988.

[3] C. Avendano and J.-M. Jot, "Frequency domain techniques for stereo to multichannel upmix," in *AES International Conference 22*, Jun. 2002.

[4] ——, "A frequency-domain approach to multichannel upmix," *J. Audio Eng. Soc*, vol. 52, no. 7/8, pp. 740–749, July 2004.

[5] M. Goto, "Active music listening interfaces based on signal processing," in *Proc. IEEE ICASSP*, 2007, pp. 1441–1444.

[6] S. Marchand, B. Mansencal, and L. Girin, "Interactive music with active audio CDs," in *Proc. CMMR*, 2010, pp. 73–74.

[7] J. Mouba and S. Marchand, "A source localization/separation/respatialization system based on unsupervised classification of interaural cues," in *Proc. DAFx*, Sep. 2006, pp. 1–6.

[8] H.-O. Oh, Y.-W. Jung, A. Favrot, and C. Faller, "Enhancing stereo audio with remix capability," in *AES Convention 129*, Nov. 2010.

[9] 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.

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

[11] 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.

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

[13] J. Ganseman, P. Scheunders, G. J. Mysore, and J. S. Abel, "Source separation by score synthesis," in *Proc. ICMC*, 2010, pp. 462–465.

[14] R. Hennequin, B. David, and R. Badeau, "Score informed audio source separation using a parametric model of non-negative spectrogram," in *Proc. IEEE ICASSP*, 2011, pp. 45–48.

[15] S. Ewert and M. Müller, "Using score-informed constraints for NMF-based source separation," in *Proc. IEEE ICASSP*, 2012, pp. 129–132.

[16] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, pp. 788–791, Oct. 1999.

[17] C. Faller and F. Baumgarte, "Efficient representation of spatial audio using perceptual parametrization," in *Proc. IEEE WASPAA*, 2001, pp. 199–202.

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

[19] 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.

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

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

[22] S. Marchand and D. Fourer, "Breaking the bounds: Introducing informed spectral analysis," in *Proc. DAFx*, Sep. 2010, pp. 1–8.

[23] 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.

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

[25] M. Parvaix and L. Girin, "Informed source separation of linear instantaneous under-determined audio mixtures by source index embedding," *IEEE Trans. Audio Speech Lang. Process.*, vol. 19, no. 6, pp. 1721–1733, Aug. 2011.

[26] A. Liutkus *et al.*, "Informed source separation through spectrogram coding and data embedding," *Signal Process.*, vol. 92, no. 8, pp. 1937–1949, Aug. 2012.

[27] N. Sturmel and L. Daudet, "Iterative phase reconstruction of Wiener filtered signals," in *Proc. IEEE ICASSP*, 2012, pp. 101–104.

[28] A. Liutkus *et al.*, "Informed audio source separation: A comparative study," in *Proc. EUSIPCO*, 2012, pp. 2397–2401.

[29] C. Falch, L. Terentiev, and J. Herre, "Spatial audio object coding with enhanced audio object separation," in *Proc. DAFx*, Sep. 2010, pp. 1–7.

[30] E. Schuijers *et al.*, "Low complexity parametric stereo coding," in *AES Convention 116*, May 2004.

[31] ISO/IEC, *Information technology — Generic coding of moving pictures and associated audio information — Part 7: Advanced Audio Coding (AAC)*, Jan. 2006, ISO/IEC 13818-7:2006.

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

[33] 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.

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

[35] "QUASI database — a musical audio signal database for source separation," http://www.tsi.telecom-paristech.fr/aao/?p=605, Mar. 2012.

[36] "Freeware Advanced Audio Coder (FAAC)," http://sourceforge.net/projects/faac/, version 1.28.

[37] ISO/IEC, *Information technology — Coding of audio-visual objects — Part 3: Audio*, Aug. 2009, ISO/IEC 14496-3:2009.

[38] ETSI, *Digital cellular telecommunications system (Phase 2+); Universal Mobile Telecommunications System (UMTS); LTE; General audio codec audio processing functions; Enhanced aacPlus general audio codec; Floating-point ANSI-C code (3GPP TS 26.410 version 10.0.0 Release 10)*, Apr. 2011, ETSI TS 126 410 V10.0.0.