HAL
open science

Informed Audio Source Separation Using Linearly Constrained Spatial Filters

Stanislaw Gorlow, Sylvain Marchand

To cite this version:

Stanislaw Gorlow, Sylvain Marchand. Informed Audio Source Separation Using Linearly Constrained Spatial Filters. IEEE Transactions on Audio, Speech and Language Processing, 2013, 21 (1), pp.3-13. 10.1109/TASL.2012.2208629 . hal-00725428

HAL Id: hal-00725428

https://hal.science/hal-00725428v1

Submitted on 22 Apr 2013

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Informed Audio Source Separation Using Linearly Constrained Spatial Filters

Stanislaw Gorlow, *Graduate Student Member, IEEE,* and Sylvain Marchand, *Senior Member, IEEE*

***Abstract*—In this work we readdress the issue of audio source separation in an informed scenario, where certain information about the sound sources is embedded into their mixture as an imperceptible watermark. In doing so, we provide a description of an improved algorithm that follows the linearly constrained minimum-variance filtering approach in the subband domain, in order to obtain perceptually better estimates of the source signals in comparison to other published approaches. Just as its predecessor, the algorithm does not impose any restrictions on the number of simultaneously active sources, neither on their spectral overlap. It rather adapts to a given signal constellation and provides the best possible estimates under given constraints in linearithmic time. The validity of the approach is demonstrated on a stereo mixture with two levels of sound complexity. It is also shown by means of both objective and subjective evaluation that the proposed algorithm outperforms a reference algorithm by at least one grade. Bearing high perceptual resemblance to the original signals at a fairly tolerable data rate of 10–20 kbps per source, the algorithm hence seems well-suited for active listening applications such as *re*-mixing or *re*-spatialization in real time.**

***Index Terms*—Array signal processing, audio quality assessment, audio watermarking, informed audio source separation.**

I. Introduction

In recent years several approaches have been proposed that address audio source separation in an "informed" scenario [1]–[3]. The reason for this new trend is the plain fact that after decades of research, "blind" or rather "semi-blind" source separation approaches to this day yield unsatisfactory quality with regard to what may be considered as "professional" audio applications, for which quality is key; for an overview and general concepts of blind approaches see [4]. Blind *speech* separation techniques, on the other hand, often rely on a speech production model and/or make specific assumptions, which generally cannot be upheld for music [5]. But what is even worse is that many sophisticated techniques are not applicable if the separation problem is *ill posed*, that is when the number of observations (channels) is smaller than the number of sources. Illposedness is yet the normal case for most music recordings, as the content is still distributed and consumed primarily in stereo format. The concept of *informed source separation* (ISS) that was first promoted by Knuth in [6] can hence be seen as a way of overcoming the limitations of blind source separation encountered in today's state-of-the-art algorithms.

The informed approach is first and foremost characterized by the *temporally and locally bounded access to the source signals*. One differentiates respectively between the processes of content creation and content consumption: The content creator provides all the necessary information for the content consumer to *decompose* the music piece into its constituent components—the source signals—in order to *recompose* the content ad libitum. Backward compatibility with conventional playback systems is further guaranteed, if the information is small enough to be inaudibly hidden in the mixture signal itself. Moreover, due to the fact that a professionally premixed version is provided by the content creator and not the source signals, intellectual property rights remain inviolate. The task of an ISS encoder is hence to extract a minimum of auxiliary information from the source signals, so that the ISS decoder can recover copies of the original signals from the mixture in high perceptual quality. The information about the sources or the signals can be embedded into the mixture signal using an audio watermarking technique, such as [7].

In the present paper we give an elaborate description for a straightforward implementation of the *underdetermined source signal recovery* (USSR) algorithm that was introduced in [3]. Generally speaking, USSR is a subband-domain beamforming technique that makes an extensive use of approximate short-time power spectral densities (STPSDs) of the source signals in order to attain an improved separation performance. In [3] it was demonstrated that a perceptually motivated approximation of the STPSDs is sufficiently precise to achieve high similarity with the original signals. In the proposed *non-iterative* variant of the algorithm, the STPSDs are also used to model the spatial correlation matrices in each point of the time-frequency (TF) plane, so as to calculate an optimum spatial filter with a desired beam response.

The idea of informing the separator with the approximate STPSDs is also found in [2]. There, however, the STPSDs are used to calculate a *generalized* Wiener filter for each source signal in each channel separately. This type of mean square error (MSE) based interference reduction takes account of the power relations between the source signals but not their spatial diversity, neither does it invert the mixing system. It is further

Manuscript received March 22, 2012; revised June 29, 2012; accepted June 30, 2012. Date of publication nulldate; date of current version nulldate. This work was partially funded by the "Agence Nationale de la Recherche" (ANR) within the scope of the DReaM project (ANR-09-CORD-006). The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Sharon Gannot.

S. Gorlow is with the Computer Science Research Laboratory of Bordeaux (LaBRI), CNRS, University of Bordeaux 1, 33405 Talence Cedex, France (e-mail: stanislaw.gorlow@labri.fr).

S. Marchand is with the Information and Communication Science and Technology Laboratory (Lab-STICC), CNRS, University of Western Brittany, 29238 Brest Cedex 3, France (e-mail: sylvain.marchand@univ-brest.fr).

Digital Object Identifier 10.1109/TASL.2012.2208629

1558–7916/$31.00 © 2012 IEEE

known that the Wiener filter minimizes the noise power at the cost of the signal of interest, which means nothing else but that the signal of interest is also attenuated for the sake of a higher signal-to-noise ratio (SNR) at the output. In the case of multiple source signals, one can equally expect the estimated source-signal spectra to be attenuated depending on whether the signal-to-interference ratio (SIR) in a TF point is high or low (for proof see Appendix). In direct consequence of this aspect, the estimated spectra of the source signals with a low SIR exhibit audibly missing spectral components, which may deteriorate the quality of the listening experience. The same effect but in a much more extreme manner was already observed with regard to the work in [1] and illustrated in [3]. The approach presented herein stands in marked contrast to the aforementioned techniques, first, because it exploits spatial diversity, and second, because it constrains the output to have a desired power level in order to overcome the issue of spectral gaps.

The rest of the paper is organized as follows: The mixture model, the problem to be solved, and the pursued approach are illustrated in Section II. The theoretical foundations for linearly constrained spatial filtering are laid in Section III, and a *power-constraining minimum-variance* (PCMV) beamformer is derived thereupon. Section IV gives a detailed description of the new non-iterative USSR algorithm, which is followed by the analysis of important performance characteristics in Section V. Objective and subjective test results are contrasted with each other in Section VI. Section VII finally concludes the paper and points out possible directions for future work.

II. Data Model, Problem Formulation, and Proposed Solution

A. Data Model

In the considered scenario, d source signals $\{s_i(n)\}_{i=1}^{d}$ are mixed into a stereo signal $\{x_1(n), x_2(n)\}$ through a linear time-invariant and memoryless system which is given by the mixing matrix $\mathbf{A} = \left[\, \mathbf{a}_1 \; \mathbf{a}_2 \; \cdots \; \mathbf{a}_d \,\right]$, with $\mathbf{a}_i = \left[\, \sin\theta_i \; \cos\theta_i \,\right]^{\mathrm{T}}$ being the panning vector associated with the ith source. Each source is positioned at the unique azimuth θ_i between fully left and fully right in the stereo sound field as illustrated in Fig. 1. The placement can be chosen either arbitrarily or following some common mixing rules. The signal power is considered inherent to the source signals and is not explicitly modeled. Furthermore, owing to the trigonometric identity

$$\|\mathbf{a}(\theta)\|^2 = \sin^2\theta + \cos^2\theta = 1, \tag{1}$$

the sound power level is kept constant across the two output channels. By rewriting the source signals in vector form as $\mathbf{s}(n) = \left[\, s_1(n) \; s_2(n) \; \cdots \; s_d(n) \,\right]^{\mathrm{T}}$ and by doing the same for the observed system output $\mathbf{x}(n) = \left[\, x_1(n) \; x_2(n) \,\right]^{\mathrm{T}}$ respectively, the latter is put in relation to $\mathbf{s}(n)$ by the mixture model

$$\mathbf{x}(n) = \sum_{i=1}^{d} \mathbf{a}_i s_i(n) = \mathbf{A}\mathbf{s}(n). \tag{2}$$

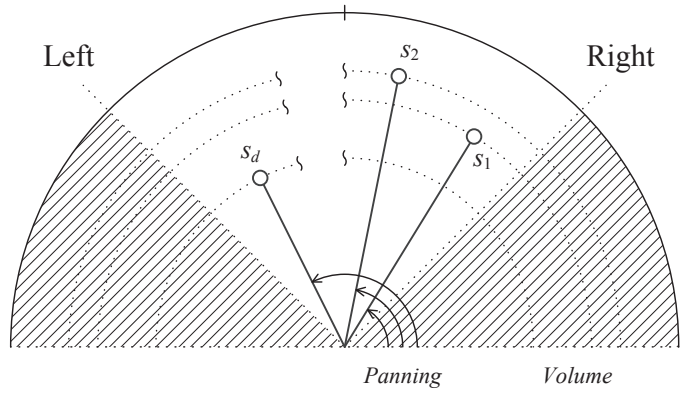

Fig. 1. Modeling of monophonic sources in the stereophonic sound field using the parameters *panning* (azimuth) and *volume* (radius). A sound source is considered unique, if the associated panning angle is unique within the considered range.

B. Problem Formulation

What we seek is a function f which transforms the mixture signal $\mathbf{x}(n)$ into an approximate source signal $\hat{s}_i(n)$ with the knowledge of the model parameters $\boldsymbol{\theta} = \left[\, \theta_1 \; \theta_2 \; \cdots \; \theta_d \,\right]^{\mathrm{T}}$ and a measurable signal characteristic $\{\phi_i(n)\}_{i=1,\ldots,d}$. The latter shall be deemed as perceptually relevant, where

$$\phi\colon \{\mathbb{R}, \mathbb{C}\} \to \mathbb{R}, s_i(n) \mapsto \phi_i(n) = \phi\left(s_i(n)\right). \tag{3}$$

Postulating the preservation of $\phi_i(n)$ in $\hat{s}_i(n)$, the problem to be solved is formulated as follows: Given $\mathbf{x}(n)$, θ_i, and $\phi_i(n)$, find

$$\hat{s}_i(n) = f\left(\mathbf{x}(n), \theta_i, \phi_i(n)\right) \tag{4}$$

such that

$$\phi\left(\hat{s}_i(n)\right) = \phi_i(n) \tag{5}$$

for $i = 1, 2, \ldots, d$.

C. Proposed Solution

In order that the source signals in (2) show a better disjoint orthogonality [8] in comparison with the waveform domain, the mixture signal $\mathbf{x}(n)$ is mapped onto an adequate time-frequency representation (TFR). The transformed mixture can then be expressed in terms of subband signals as

$$\mathbf{x}_k(m) = \sum_{i=1}^{d} \mathbf{a}_i s_{ik}(m) = \mathbf{A}\mathbf{s}_k(m), \tag{6}$$

where k represents the subband index and m denotes the index of the preferably complex-valued[1] time series in that particular subband. Each mixture signal $\mathbf{x}_k(m)$ is then decomposed into its constituent parts by use of linear spatial filtering according to

$$\hat{s}_{ik}(m) = \sum_{c=1}^{2} w_{ick} x_{ck}(m) = \mathbf{w}_{ik}^{\mathrm{T}} \mathbf{x}_k(m), \tag{7}$$

where $\mathbf{w}_{ik} = \left[\, w_{i1k} \; w_{i2k} \,\right]^{\mathrm{T}}$ is the spatial filter, or beamformer, that provides an estimate $\hat{s}_{ik}(m)$ for the ith signal component in the kth subband based on the observed mixture signal, and

[1]The use of a real-valued filter bank may cause aliasing artifacts.

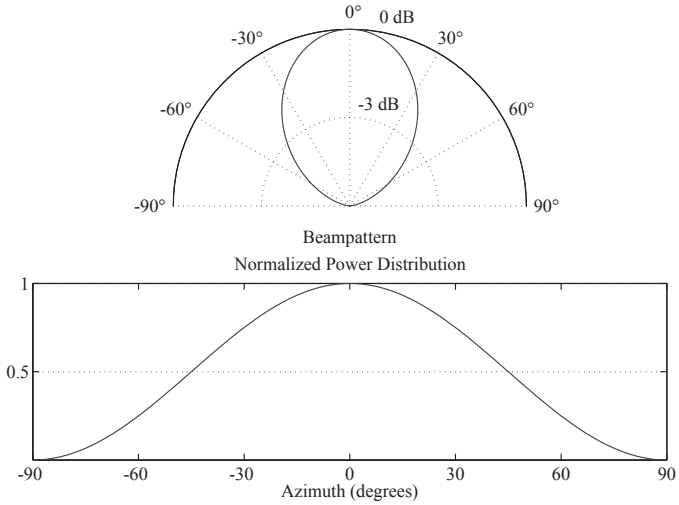

Fig. 2. Beampattern and power distribution of the stereo beamformer. The beam is gain adjusted and directed such that the signal of interest is preserved and either one interferer is suppressed or, in case of multiple interferers, the total power of all interferers is minimized.

$\mathbf{w}_{ik}^{\mathrm{T}}$ is its transpose. Since the mixing system that we seek to invert is real-valued, so is the beamformer. From a geometrical point of view, the beam in Fig. 2 is steered and amplified (or attenuated), so that the signal component in the direction of θ_i is preserved, while the contribution from the interfering sources is canceled out or at least minimized. In the latter case, the beamformer shall be constrained to adjust the power level of the output signal to the instantaneous power level of the original signal—$|s_{ik}(m)|^2$—in the respective subband. This is formally achieved by choosing ϕ in (3) as the squared magnitude. The filtered subband signals are then recombined into an isolated version of the original source signal $\hat{s}_i(n)$.

III. Linearly Constrained Spatial Filtering

A. Preliminaries

Before going into detail about linearly constrained spatial filtering, let us recall two typical beamforming quantities that will be used in this paper first (see also [9]). The indices i and k are omitted for simplicity.

1) The azimuth-dependent *beam response* or *gain* is defined as
$$g(\theta) \triangleq \mathbf{w}^{\mathrm{T}}\mathbf{a}(\theta), \tag{8}$$
and the so-called *beampattern* is the magnitude-squared beam response.
2) The *spatial power spectrum*
$$P(\theta) \triangleq \mathbf{w}^{\mathrm{T}}(\theta)\mathbf{R}_{\mathbf{x}}\mathbf{w}(\theta) \tag{9}$$
is a measure for the mean total sound power received from the direction of θ and $\mathbf{R}_{\mathbf{x}}$ is the spatial correlation matrix of the two-channel mixture vector $\mathbf{x}$.

B. Signal Model

Following classical literature on statistical signal processing, the source signals are modeled as zero-mean Gaussian random processes that are mutually independent and non-stationary. Joint wide-sense stationarity can nonetheless be presumed for the duration of a sufficiently short time segment. The short-time power spectral density (STPSD or PSD for short) may then be used as a measure for how the mean signal power or variance distributes with time and frequency (see the Wiener–Khinchin convergence theorem).

C. Well-Posed Case ($1 \leqslant d \leqslant 2$)

The term "well posed" shall characterize the case when the number of active sound sources is at least one but not larger than the number of channels, which is two. In such a case, it exists one exact solution.

1) Unity-Gain Filter: Let us suppose that the mixture signal $\mathbf{x}_k(m)$ constitutes only one directional source signal,
$$\mathbf{x}_k(m) = \mathbf{a}_i s_{ik}(m). \tag{10}$$
By comparing (1) with (8) under the unity-gain constraint
$$g_k(\theta_i) \stackrel{!}{=} 1, \tag{11}$$
it becomes evident that the source signal component $s_{ik}(m)$ can be extracted from the mixture $\mathbf{x}_k(m)$ by setting
$$\mathbf{w}_{ik} = \mathbf{a}_i, \tag{12}$$
so that
$$s_{ik}(m) = \mathbf{a}_i^{\mathrm{T}}\mathbf{x}_k(m). \tag{13}$$

2) Zero-Forcing Filter: Now let two sources contribute to the mixture $\mathbf{x}_k(m)$ simultaneously:
$$\mathbf{x}_k(m) = \mathbf{a}_i s_{ik}(m) + \mathbf{a}_l s_{lk}(m), \tag{14}$$
where $s_{ik}(m)$ is the signal of interest and $s_{lk}(m)$ shall denote the jammer respectively. By enforcing identity as in (11) for $s_{ik}(m)$ and full cancellation of $s_{lk}(m)$ by
$$g_k(\theta_l) \stackrel{!}{=} 0, \tag{15}$$
the sought-after weight vector calculates from
$$\mathbf{w}_{ik} = \left(\mathbf{A}^{-1}\right)^{\mathrm{T}}\mathbf{g}_{ik}, \tag{16}$$
where $\mathbf{g}_{ik} = \left[\begin{smallmatrix}1 & 0\end{smallmatrix}\right]^{\mathrm{T}}$ is the gain vector and $\mathbf{A}^{-1}$ is the inverse of the corresponding mixing matrix $\mathbf{A} = \left[\begin{smallmatrix}\mathbf{a}_i & \mathbf{a}_l\end{smallmatrix}\right]$. Applying the above procedure for both sources yields the separation matrix $\mathbf{W}_k = \left[\begin{smallmatrix}\mathbf{w}_{ik} & \mathbf{w}_{lk}\end{smallmatrix}\right] = \left(\mathbf{A}^{-1}\right)^{\mathrm{T}}$, and the source signals $\mathbf{s}_k(m) = \left[\begin{smallmatrix}s_{ik}(m) & s_{lk}(m)\end{smallmatrix}\right]^{\mathrm{T}}$ are obtained less surprisingly from
$$\mathbf{s}_k(m) = \mathbf{A}^{-1}\mathbf{x}_k(m). \tag{17}$$

D. Ill-Posed Case ($d > 2$)

The term "ill posed", as opposed to "well posed", shall be used as a synonym for the case when a unique solution to the source separation problem does not exist, that is when the mixture is composed of more than two source signals. An optimum solution can be found instead by means of *linearly constrained minimum-variance* (LCMV) filtering [10]. What is considered as the jammer then is the sum of *all* interfering source signals. In consequence, the mixture signal $\mathbf{x}_k(m)$ is

equally modeled in terms of two components: a unidirectional signal of interest and a multidirectional jammer, i.e.

$$\mathbf{x}_k(m) = \mathbf{a}_i s_{ik}(m) + \mathbf{r}_k(m), \tag{18}$$

where $\mathbf{r}_k(m) = \sum_{l,l\neq i} \mathbf{a}_l s_{lk}(m)$. An estimate $\hat{s}_{ik}(m)$ for the signal of interest $s_{ik}(m)$ is found by minimizing the mean jammer power, or equivalently the mean beamformer output power, along the direction of the signal of interest

$$P_k(\theta_i) = \mathbf{w}_{ik}^{\mathrm{T}}(m)\mathbf{R}_{\mathbf{x}_k}(m)\mathbf{w}_{ik}(m), \tag{19}$$

subject to identity with respect to a given power level $\phi_{ik}(m)$. In other words, we seek after the weight vector that solves the quadratic optimization problem

$$\begin{aligned} \mathbf{w}_{ik\mathrm{o}}(m) &= \underset{\mathbf{w}_{ik}(m)}{\arg\min}\, P_k(\theta_i) \\ \text{s.t. } g_k(\theta_i) &\overset{!}{=} \sqrt{\phi_{ik}(m)\mathbf{a}_i^{\mathrm{T}}\mathbf{R}_{\mathbf{x}}^{-1}\mathbf{a}_i}. \end{aligned} \tag{20}$$

The desired solution to the above problem is found, e.g., by use of the method of Lagrange multipliers, which yields

$$\mathbf{w}_{ik\mathrm{o}}(m) = \mathbf{R}_{\mathbf{x}_k}^{-1}(m)\mathbf{a}_i\sqrt{\frac{\phi_{ik}(m)}{\mathbf{a}_i^{\mathrm{T}}\mathbf{R}_{\mathbf{x}_k}^{-1}(m)\mathbf{a}_i}}. \tag{21}$$

When applied to the mixture signal $\mathbf{x}_k(m)$, the derived filter will narrow the lobe of the jammer power spectrum and the leakage from the interfering sources into the estimated signal of interest, hence, will be reduced. Due to the power constraint, it is furthermore ensured that the power level of the estimate will match with the desired power level in every point of the time-frequency plane. This can be easily verified by plugging (21) in into (19).

E. Noisy Case

In the case where watermarking forms a part of the encoder, the mixture model in (6) can be extended by a noise term in the following way:

$$\mathbf{x}_k^{\mathrm{w}}(m) = \mathbf{A}\mathbf{s}_k(m) + \mathbf{n}_k(m), \tag{22}$$

where $\mathbf{x}_k^{\mathrm{w}}(m)$ is the watermarked mixture signal and $\mathbf{n}_k(m)$ is an additive noise component in the respective TF point. Due to the fact that the employed watermarking technique is based on the on-frequency masking phenomenon, the noise term can be assumed to be *collinear* with the noise-free mixture signal, which results in the following relation:

$$\mathbf{x}_k^{\mathrm{w}}(m) = [1 + \eta_k(m)]\,\mathbf{x}_k(m), \tag{23}$$

with $\mathbf{n}_k(m) = \eta_k(m)\mathbf{x}_k(m)$, and $\eta_k(m)$ being a scalar that represents the corruption due to the watermark. From (23) it is evident that the estimate $\hat{s}_{ik}(m) = \mathbf{w}_{ik}^{\mathrm{T}}\mathbf{x}_k^{\mathrm{w}}(m)$ needs to be rectified by $[1 + \eta_k(m)]^{-1}$, so as to compensate for the watermark. In the general case, however, $1 + \eta_k(m)$ will not be known. We can yet give an estimate for the deviation of the magnitude with respect to a (known) noise-free power level $\phi_{ik}(m)$, which is

$$|1 + \eta_k(m)| = \frac{|\hat{s}_{ik}(m)|}{\sqrt{\phi_{ik}(m)}}. \tag{24}$$

This a posteriori estimate for magnitude distortion can then be used to partly compensate for errors due to watermarking.

F. General Remarks

If the constraints from (11) and (15) are imposed on the beamformer in (20) at the same time, as in [3], the obtained solution folds up to the expression in (16), simply because no degrees of freedom are left with regard to the number of weight coefficients to minimize the jammer power. This has as consequence that the estimator is definitely suboptimal in the case of multiple interferers: Canceling out just one of the interfering sources analogously to (16) often enough leaves a strong residual, which is further amplified by the filter.

Another very popular yet unconstrained solution, which is the best solution in the MSE sense, can be obtained by leaving out the identity constraint in (20). The corresponding weight coefficients of such a "spatial" Wiener filter, alias minimum mean-square error (MMSE) beamformer, are [11, ch. 2]

$$\mathbf{w}_{ik\mathrm{o}}^{\mathrm{MMSE}}(m) = \mathbf{R}_{\mathbf{x}_k}^{-1}(m)\mathbf{p}_{ik}(m), \tag{25}$$

where $\mathbf{p}_{ik}(m) = \mathrm{E}\left[\mathbf{x}_k(m)s_{ik}^*(m)\right]$ is the cross-correlation signal between the mixture $\mathbf{x}_k(m)$ and the complex conjugate of the signal of interest $s_{ik}^*(m)$. E further denotes the statistical expectation operator. Given that $\mathbf{p}_{ik}(m) = \mathbf{a}_i\sigma_{ik}^2(m)$, where σ^2 is the variance, it then follows that $\mathbf{w}_{ik\mathrm{o}}(m) \sim \mathbf{w}_{ik\mathrm{o}}^{\mathrm{MMSE}}(m)$, and from this one can infer that the two beams have the same look direction but different gains: The beamformer from (21) adapts the gain in order to conform with the power constraint, whereas the MMSE beamformer will likewise power down the output signal for the sake of a lower MSE. This issue is addressed in the Appendix.

IV. The Algorithm

A. System Overview

A schematic overview of the algorithm that was given the name "Underdetermined Source Signal Recovery" (USSR) is shown in Fig. 3. Although some naming differences can be found with regard to [3], the processing steps are essentially identical.

1) Encoder: The source signals $\{s_i(n)\}_{i=1}^d$ are blockwise time-frequency mapped by means of the short-time Fourier transform (STFT) as stated by the formula below:

$$s_{ik}(m) = \sum_{n=0}^{L-1} s_i(mM + n)w(n)e^{-j2\pi/Nkn}, \tag{26}$$

$0 \leqslant k < N$, where m is the segment index, L is the size of the window $w(n)$, M is the window shift size ($0 < M \leqslant L$), j is the imaginary unit, and N is the transform length ($L \leqslant N$). The instantaneous power signal $\phi_{ik}(m)$ is calculated next according to

$$\phi_{ik}(m) = |s_{ik}(m)|^2, \tag{27}$$

where $\boldsymbol{\phi}_i(m) = \left[\,\phi_{i,0}(m)\ \phi_{i,1}(m)\ \cdots\ \phi_{i,N-1}(m)\,\right]$ is equally the time-varying PSD of the ith source signal. All d STPSDs are approximated and quantized on a double-logarithmic scale (see Section IV-B) and afterwards hidden together with the model parameters $\boldsymbol{\theta}$ in the waveform signal $\mathbf{x}(n)$ as an imperceptible watermark. Finally, the watermarked mixture signal $\mathbf{x}^{\mathrm{w}}(n)$ is communicated to the decoder.

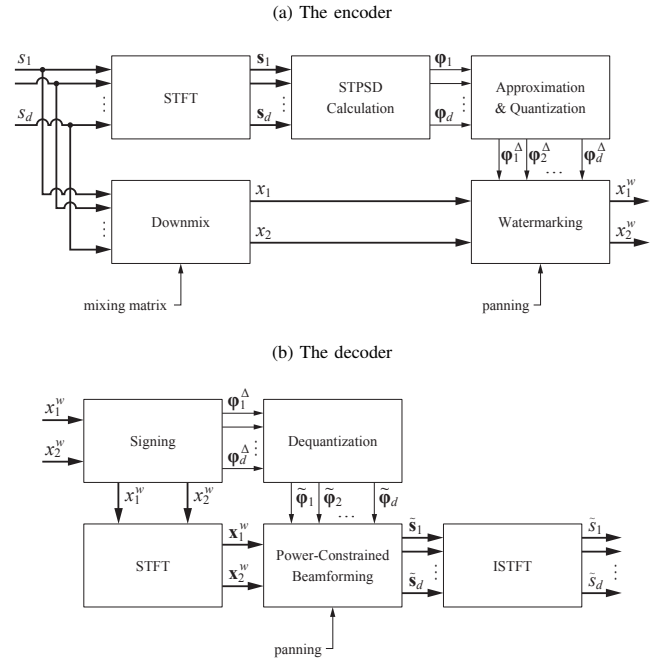

Fig. 3. Functional block diagram of the encoder and the decoder. The source signals are transmitted to the decoder in a watermarked downmix and recovered from the latter using approximate power spectral densities and the panning angles.

2) Decoder: The incoming signal $\mathbf{x}^{\mathrm{w}}(n)$ is first "signed", meaning that the watermark is extracted, and then channelwise TF transformed to obtain the subband signals $\{\mathbf{x}_k^{\mathrm{w}}(m)\}_{k=0}^{N-1}$. The latter are then processed by the separator kernel which avails himself of the dequantized STPSDs $\tilde{\boldsymbol{\phi}}_i(m)$ (see Section IV-C) and the model parameters $\boldsymbol{\theta}$; the exact procedure is given in Section IV-D. The filtered signal spectra $\{\tilde{\mathbf{s}}_i(m)\}_{i=1}^{d}$, where $\tilde{\mathbf{s}}_i(m) = \left[\, \tilde{s}_{i,0}(m) \; \tilde{s}_{i,1}(m) \; \cdots \; \tilde{s}_{i,N-1}(m) \,\right]$, are then transformed back to the waveform domain using the inverse STFT (ISTFT) [12, ch. 10].

B. Approximation and Quantization

A significant reduction of side information can be achieved in two ways: first, by reducing the frequency resolution of the PSDs $\boldsymbol{\phi}_i(m)$ in approximation of the critical bands [13], and second, by quantizing the PSD values $\phi_{ik}(m)$ with a step size equal to some value Δ, which is put in relation to an appropriate psychoacoustic criterion.

The peripheral auditory system is usually modeled as a bank of overlapping bandwidth filters, the auditory filters, which possess an equivalent rectangular bandwidth (ERB). The scale that relates the center frequency of auditory filters to units of the ERB is the ERB-rate scale. Using the ERB-rate function in [14] we can define a relation between the frequency index k and the critical-band index z_k by

$$z_k \triangleq \lfloor 21.4 \log_{10}(4.37 f_s/Nk + 1) \rfloor, \tag{28}$$

where $\lfloor \cdot \rfloor$ is the floor function and f_s is the sampling frequency in kHz. The zth critical-band value of the approximate PSDs is then calculated as the arithmetic mean between $\mathrm{lb}(z) = \inf\{k\colon z_k = z\}$ and $\mathrm{ub}(z) = \sup\{k\colon z_k = z\}$ according to

$$\bar{\phi}_{iz}(m) = \frac{1}{\mathrm{ub}(z) - \mathrm{lb}(z) + 1} \sum_{k=\mathrm{lb}(z)}^{\mathrm{ub}(z)} \phi_{ik}(m). \tag{29}$$

Furthermore, under the assumption that the the minimum just-noticeable-difference level and so the maximum allowed quantization error is 1 dB [13], the quantization step size Δ is chosen as 2 dB, and the irrelevancy-reduced PSD values are obtained from the uniform quantizer

$$\phi_{iz}^{\Delta}(m) = \left\lfloor 5 \log_{10} \bar{\phi}_{iz}(m) \right\rceil, \tag{30}$$

where $\lfloor \cdot \rceil$ denotes the round-to-nearest rounding function. The panning angles θ_i are simply rounded to the nearest integer value.

C. Dequantization

As the beamforming is carried out based on the availability of short-time PSDs, the signed PSD values are converted back into linear PSD values and extrapolated to the resolution of the STFT. This operation, which is labeled as "dequantization" in Fig. 3b, is performed by

$$\tilde{\phi}_{ik}(m) = 10^{\phi_{iz}^{\Delta}(m)/5} \qquad \forall k\colon z_k = z. \tag{31}$$

D. Power-Constrained Beamforming

In analogy with Section III, we differentiate between (over-) *determined* and *underdetermined* TF regions. Determined TF regions are points in the TF plane with at least one and at most two active sources, whereas underdetermined regions are TF points with more than two active sources respectively. The indices of the active sources are determined by comparing the PSD values $\tilde{\phi}_{ik}(m)$ in a given TF point with the noise-floor power level. Sources for which the PSD value is greater than say -60 dB are consequently deemed as active. Determined TF regions with the same source indices may beyond be grouped into clusters. The noise cluster would hence be formed by TF regions without any sources. Each signal component of an active source in a given TF point or cluster is then filtered out of the mixture according to one of the three cases listed below.

1) One active source: If one single source was detected in the respective TF region, the signal component is calculated as

$$\hat{s}_{ik}(m) = \mathbf{a}_i^{\mathrm{T}} \mathbf{x}_k^{\mathrm{w}}(m). \tag{32}$$

2) Two active sources: If two sources were detected, each of the two signal components is obtained by

$$\hat{s}_{ik}(m) = \left[\, 1 \; 0 \,\right] \left[\, \mathbf{a}_i \; \mathbf{a}_l \,\right]^{-1} \mathbf{x}_k^{\mathrm{w}}(m), \tag{33}$$

with interchanging the indices.

3) More than two active sources: If more than two sources were detected in the same TF point, the beamformer from (21) is steered for each of the signal components separately, so that each component is estimated as

$$\hat{s}_{ik}(m) = \sqrt{\frac{\tilde{\phi}_{ik}(m)}{\mathbf{a}_i^{\mathrm{T}}\tilde{\mathbf{R}}_{\mathbf{x}_k}^{-1}(m)\mathbf{a}_i}}\,\mathbf{a}_i^{\mathrm{T}}\tilde{\mathbf{R}}_{\mathbf{x}_k}^{-1}(m)\mathbf{x}_k^{\mathrm{w}}(m), \tag{34}$$

where the output correlation matrix is approximated as

$$\tilde{\mathbf{R}}_{\mathbf{x}_k}(m) = \sum_{i=1}^{d} \tilde{\phi}_{ik}(m)\mathbf{a}_i\mathbf{a}_i^{\mathrm{T}}. \tag{35}$$

The beamformer will hence seek to spatially decorrelate the ith signal component from the rest, while adjusting the signal amplitude to the desired level at the same time.

In the case where the amount of hidden information is quite large, so that the watermarking noise cannot be neglected, the estimated signal components are in addition power adjusted, as stated by

$$\tilde{s}_{ik}(m) = \frac{\hat{s}_{ik}(m)}{|\hat{s}_{ik}(m)|}\sqrt{\tilde{\phi}_{ik}(m)}. \tag{36}$$

V. Performance Characteristics

A. Algorithmic Delay

The algorithmic delay of the USSR algorithm is determined by the framing and overlap delay of the STFT and its inverse. By using a 2048-length symmetric window with 50-% overlap between consecutive frames, the algorithmic delay amounts to 2047 samples. This value corresponds to 46.4 ms at a sampling rate of 44.1 kHz.

B. Computational Complexity

The filtering performed by the USSR algorithm depends on the frequency content of source signals and thus varies from one mixture to another. It is therefore convenient to analyze the run-time complexity for the worst-case scenario in terms of "big O" notation. Moreover, we seek to establish a relationship between the running time and the following input parameters: the number of sources d, the number of frequency bands Z, and the transform length N, where $d < Z < N$. We further postulate that all arithmetic operations that we want to count require exactly one unit of time to execute. The results of the analysis are summarized in Table I. The figures reveal that the decoder has a complexity that is comparable to the complexity of the encoder. Evidently, the execution time is dominated by the d-fold STFT and its inverse.

C. Information Rate

The information rate of the USSR algorithm is measured by the number of bits that are communicated to the decoder per time frame excluding the mixture signal. These comprise the PSD values and the panning angles. The panning angles have a payload of 7 bit times the number of sources and are signaled to the decoder once at the beginning of the transmission. The PSD values are represented as 6-bit unsigned integers. Table II provides a brief overview of the capacities that are necessary to store the PSD values for various numbers of frequency bands. They were calculated according to

$$\text{bitrate} = \frac{f_s}{M}\cdot 6\cdot Z\cdot d, \qquad \text{with } M = L/2. \tag{37}$$

The information rate is therefore varied upon a subdivision the ERB by an integer factor. As a rule of thumb, the finer is the frequency resolution the higher is the observed quality of the spectrum estimates. On the other hand, the larger is the number of source signals, the higher is their spectral overlap, and the finer is to be chosen the frequency resolution to have a quality that is comparable to a sparser configuration. Whatever the case, the numbers in Table II can be drawn upon to make an estimate for the information rate of the USSR algorithm, as the rate of the panning angles is comparably negligible.[2]

To reduce the amount of side information even more, one can exploit the correlation of quantized PSD values between adjacent TF points. This can be achieved, e.g., by calculating the difference between two consecutive PSD values either in time or frequency direction and by coding the difference signal based on its entropy. This principle is well known as first-order differential pulse-code modulation (DPCM).

We have validated the concept by modeling the probability distribution of the input symbols based on the occurrence of

TABLE I
Run-time complexity of the USSR algorithm as a function of the number of sources d, the number of frequency bands Z, and the transform length N

Subroutine	Arithmetic operations in units of time
STFT	$\mathrm{O}(dN\log N)$
STPSD Calculation	$\mathrm{O}(dN)$
Approximation & Quantization	$\mathrm{O}(dN)$
Downmix	$\mathrm{O}(dN)$
$T_{\text{enc}}(d, Z, N) = \mathrm{O}(dN\log N)$	
STFT	$\mathrm{O}(N\log N)$
Dequantization	$\mathrm{O}(dZ)$
Power-Constrained Beamforming	$\mathrm{O}(dN)$
ISTFT	$\mathrm{O}(dN\log N)$
$T_{\text{dec}}(d, Z, N) = \mathrm{O}(dN\log N)$	

TABLE II
Information rate of PSD values (left) and PSD difference values using TFDPCM and Huffman coding (right) at 44.1 kHz sampling rate and 16-kHz cutoff

ERB subdivision factor	Number of frequency bands	Information rate in kbps per source
–	39	10.1 / **5.88**
2	76	19.6 / **11.5**
3	108	28.0 / **16.3**
4	136	35.2 / **20.6**
5	163	42.2 / **24.6**
⋮	⋮	⋮

[2]In the particular case of an instantaneous mixture, the mixing coefficients may just as well be estimated from the mixture signal itself using the algorithm in [15]. The transmission of the panning angles can then be entirely omitted.

each possible difference value in a training set. It was observed that the difference signal both in time and frequency direction has a Laplace(μ, b) distribution with $\hat{\mu} \approx -0.2$ and $\hat{b} \approx 2$. We have thereupon derived a Huffman codebook [16] from the estimated input probability distribution with a mean codeword length of 3.5 bit. This corresponds to a compression ratio of 1.7:1, which means that almost twice as many source signals can now be extracted from the mixture for the same amount of side information. The corresponding bitrates are listed in Table II in boldface.

VI. Quality Assessment

A. Algorithms Under Test

The following two algorithms were compared against each other: the USSR algorithm and an in-house implementation of the algorithm described in [1]. These algorithms were already compared in [3], but this time the USSR algorithm is served in three different flavors. The tested algorithms hence are:

- ISSA — The reference ISS algorithm from [1].
- USSR-A — The *iterative* USSR algorithm [3] with *pairwise extraction*.
- USSR-B — The proposed *non-iterative* USSR algorithm.
- USSR-C — USSR-B in combination with the *unconstrained* Wiener–Hopf solution from (25).

B. Test Items

We had selected two pieces from different music genres: a 5-track hip-hop mixture and a more complex 7-track pop-rock mixture. The hip-hop piece was DJ Vadim's "The Terrorist". It is composed of a leading vocal, a synthesizer in the bassline, and a percussion section that includes a kick, a snare, and a hi hat. Phoenix's "Lisztomania" was chosen from within the pop-rock genre. It has a bass guitar together with drums forming the rhythmic section, several guitars in the harmonic section, a vocal melody, and a keyboard to create a sustained pad for the song. The signals used in the test were 30-s long monophonic excerpts from the multitrack masters. Their spatial placement, which was aligned with the commercial publications, is given in Table III.

C. Test Conditions

To exclude a performance bias due to different TFRs, all four algorithms were implemented using the STFT. The STFT was realized as a 2048-point fast Fourier transform (FFT) with a Kaiser–Bessel derived window of the same length and a 50-% overlap between succeeding frames. The sampling rate was set to 44.1 kHz. The effective data rate of the ISSA algorithm was 86.1 kbps for the 5-track mixture and 108 kbps for the 7-track mixture, respectively. The USSR algorithm was tuned in such a way that the raw bitrates were approximately the same: 93.4 and 103 kbps. In addition, the same audio watermarking technique [7] was used in all four cases.

TABLE III
Panning used for the two music pieces

Track name	Panning
Acapella	6.7 % right
Bass	20 % left
Hi Hat	29 % left
Kick	6.7 % left
Snare	centered
"The Terrorist" by DJ Vadim	
Bass	1.6 % right
Beat	4.4 % left
Cocotte	41 % right
Guitar 1	9.3 % left
Guitar 3	76 % left
Key	9.6 % right
Vox	4.0 % right
"Lisztomania" by Phoenix	

D. Objective Performance Metrics

The following four metrics were used in order to assess the audio quality of the algorithms: the signal-to-interference ratio (SIR), the so-termed "target-related" perceptual score (TPS), a frequency-weighted signal-to-noise ratio (SNRF) [17], and the "auditory" bandwidth as the counterpart of the "articulatory" bandwidth [17]. The first two metrics were computed with the PEASS toolkit [18]. The SNRF was redefined in the following manner:

$$\mathrm{SNRF}_i(m) \triangleq \frac{1}{Z}\sum_{z=1}^{Z} 10\log_{10}\frac{\bar{\phi}_{iz}(m)}{\bar{\phi}_{iz,n}(m)}, \tag{38}$$

where $Z = 39$, z is the ERB-scale index, $\bar{\phi}_{iz}(m)$ is the average power of the reference signal, and $\bar{\phi}_{iz,n}(m)$ is the corresponding average noise power,

$$\bar{\phi}_{iz,n}(m) = \sum_{k=\mathrm{lb}(z)}^{\mathrm{ub}(z)} \frac{\left[\min\left(|\tilde{s}_{ik}(m)| - |s_{ik}(m)|, 0\right)\right]^2}{\mathrm{ub}(z) - \mathrm{lb}(z) + 1}. \tag{39}$$

The noise signal is calculated in such a way as to accentuate the subjective effect of spectral gaps on audio quality, as only *lacking* signal components are taken into account. We further used a time resolution of 23.2 ms for both the SNRF_i and the bandwidth measure. The final score was obtained by taking the average over all time segments.

E. Mean Opinion Scores

As a supplement to the objective metrics, a multi-stimulus test with hidden reference and anchor (MUSHRA) [19] has been administered, so as to obtain a set of *subjective* scores. The latter were intended to help verify consistency between objective performance metrics and human perception. The test was carried out in the audiovisual production facilities within the University of Western Brittany. Sennheiser HD 650 headphones and MOTU's UltraLite-mk3 Hybrid audio interface were used during the test for sound reproduction. The gain of the preamplifier stage was adjusted to a reference listening level of -20 dB below the clipping level of a digital

tape recording. All test signals were shortened to 20 s at the longest and the anchor was a 3.5-kHz lowpass filtered version of the sum of the original source signals with a 3-dB signal-to-interference ratio and a 50-% spectral-gap rate. The anchor was altered in such a way as to show similar types of impairment as the algorithms under test. Nine audiovisual media students have taken part in the test. They were instructed to score the stimuli according to the continuous quality scale by judging their degree of preference for one type of artifact versus some other type.

F. Test Results

The results of quality assessment are summarized in Figs. 4–8. Figs. 4–5 show the SIR and the TPS for each track from the two selected music excerpts. The corresponding SNRFs as well as the auditory bandwidths are depicted in Figs. 6–7. The mean opinion scores (MOSs) including the 95-% confidence intervals are plotted in Fig. 8. As it was anticipated, USSR-C exhibits the highest SIR. USSR-B shows a clear improvement over USSR-A. The SIR for ISSA is also quite high but always lower than for USSR-B and USSR-C however. The TPS is fairly consistent in all three USSR variants for the hip-hop mixture, whereas a slight tendency towards USSR-B can be observed for the pop-rock mixture. ISSA has definitely the worst TPS of all tested ISS algorithms. In respect of the SNRF, the two constrained USSR variants, A and B, show superiority over the rest. Again, this is something that could be expected, since these algorithms preserve the auditory bandwidth of the signal. Moreover, it can be seen that the number of spectral gaps is lesser with USSR-C than with ISSA, which supports the statement made in the introductory paragraph. De facto, the effect observed with USSR-C is more of a band limitation than the "arid" effect [3], and as such it produces a sound that is rather "dull" than "annoying". On the whole, the preferential tendencies of the TPS have shown to be rather consistent with the MOS. Yet, the TPS seems to overrate the audio quality by some 20–40 points, which again corresponds to 1–2 grades! In this regard, the SNRF has too proven to provide the desired tendencies, which allows the conclusion that if it was properly scaled, it might just as well serve as an objective measure for the perceived audio quality but at a much lower computational cost. With a mean score between "good" and "fair", the USSR algorithm came off as the clear winner in any of its variants. ISSA was overall graded as "bad"—but better than the anchor. A slight preference of USSR-B, which is the improved version of the algorithm, to its predecessor USSR-A could also be noted. That preference would seem to be linked to the complexity of the mixture. But above all, USSR-B was assessed to perform *significantly* better than USSR-C, which once more highlights the fact that full bandwidth is essential for a natural listening experience.

VII. Conclusion

In the present paper we have given a detailed description of a non-iterative version of the USSR algorithm that runs in linearithmic time. The fact that the spatial correlation matrices are derived directly from the panning angles and the STPSDs, and no longer from the clustered data, renders the algorithm less complex and more efficient at the same time. This was verified with both objective and subjective quality metrics.[3] The equal-power ($\sim$ loudness) constraint, which is inherent in the newly derived PCMV beamformer, guarantees that the recovered replica of the original source signals feature the same (approximated) STPSDs and are perceptually more similar in timbre to the latter.

In conclusion, the provided framework and the used notation permit the mixture signal to be processed either frame by frame or as a whole. But besides that, the proposed algorithm could be applied to convolutional and/or time-variant mixture models, which is an outlook on future work. Another area of interest is to extract spatial images of stereophonic sources, that is to separate the sources without changing their spatial location. It is also thinkable to combine multiple constraints, if the mixture signal has more than two channels.

Appendix

Based on the signal model from Section III-B, the MMSE beamformer for an arbitrary frequency subband and for the duration of an arbitrary time segment can be reformulated as

$$\begin{aligned}\mathbf{w}_{io}^{\text{MMSE}} &= \sigma_i^2 \mathbf{R}_{\mathbf{x}}^{-1} \mathbf{a}_i \\ &= \frac{\sigma_i^2}{\det \mathbf{R}_{\mathbf{x}}} \operatorname{adj} \mathbf{R}_{\mathbf{x}} \mathbf{a}_i \\ &= \frac{\sigma_i^2}{\det \mathbf{R}_{\mathbf{x}}} \sum_{l=1}^{d} \sigma_l^2 \operatorname{adj}\big(\mathbf{a}_l \mathbf{a}_l^{\mathrm{T}}\big) \mathbf{a}_i \\ &= \frac{\sigma_i^2}{\det \mathbf{R}_{\mathbf{x}}} \sum_{l=1}^{d} \varrho_{il} \sigma_l^2 \mathbf{Q} \mathbf{a}_l, \end{aligned} \tag{40}$$

where $\det \mathbf{R}_{\mathbf{x}}$ is the determinant and $\operatorname{adj} \mathbf{R}_{\mathbf{x}}$ is respectively the adjugate of $\mathbf{R}_{\mathbf{x}}$. Moreover, $\varrho_{il} = \det\left[\,\mathbf{a}_i\ \mathbf{a}_l\,\right] = \mathbf{a}_l^{\mathrm{T}} \mathbf{Q} \mathbf{a}_i$, with $\mathbf{Q} = \left[\begin{smallmatrix} 0 & 1 \\ -1 & 0 \end{smallmatrix}\right]$. $\det \mathbf{R}_{\mathbf{x}}$ further unfolds to

$$\begin{aligned}\det \mathbf{R}_{\mathbf{x}} &= \sum_{u=1}^{d} \sigma_u^2 a_{1u} \sum_{v=1}^{d} \varrho_{uv} \sigma_v^2 a_{2v} \\ &= \sum_{u=1}^{d} \sigma_u^2 \sum_{v=u}^{d} \varrho_{uv}^2 \sigma_v^2. \end{aligned} \tag{41}$$

The beamformer gain along the direction of the ith source is

$$\begin{aligned} g_{io}^{\text{MMSE}} &= \big(\mathbf{w}_{io}^{\text{MMSE}}\big)^{\mathrm{T}} \mathbf{a}_i \\ &\overset{(40)}{=} \frac{\sigma_i^2}{\det \mathbf{R}_{\mathbf{x}}} \sum_{l=1}^{d} \varrho_{il} \sigma_l^2 \underbrace{\mathbf{a}_l^{\mathrm{T}} \mathbf{Q} \mathbf{a}_i}_{\varrho_{il}} \\ &\overset{(41)}{=} \frac{\sigma_i^2 \sum_{l=1}^{d} \varrho_{il}^2 \sigma_l^2}{\sum_{u=1}^{d} \sigma_u^2 \sum_{v=u}^{d} \varrho_{uv}^2 \sigma_v^2}. \end{aligned} \tag{42}$$

In the ill-posed case, that is for $d > 2$, it can be noted that $0 \leqslant \varrho_{il}^2, \varrho_{uv}^2 < 1$, and since $\sigma_l^2, \sigma_v^2 \geqslant 0$, the following inequalities

[3]Download the sound clips from http://www.labri.fr/~gorlow/lcsf/

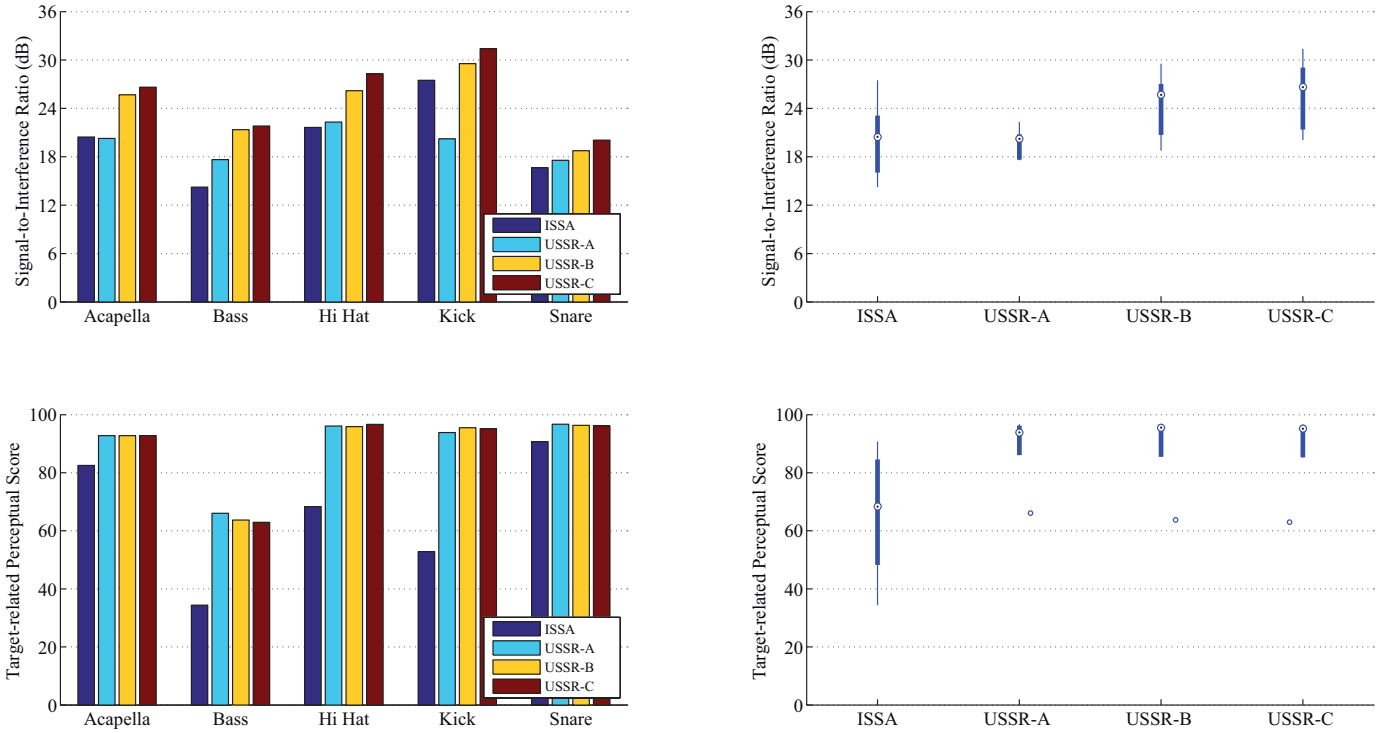

Fig. 4. Signal-to-interference ratios and target-related perceptual scores (left column) complemented by the median, the 25th and 75th percentiles, and outliers (right column) for an excerpt from "The Terrorist" by DJ Vadim.

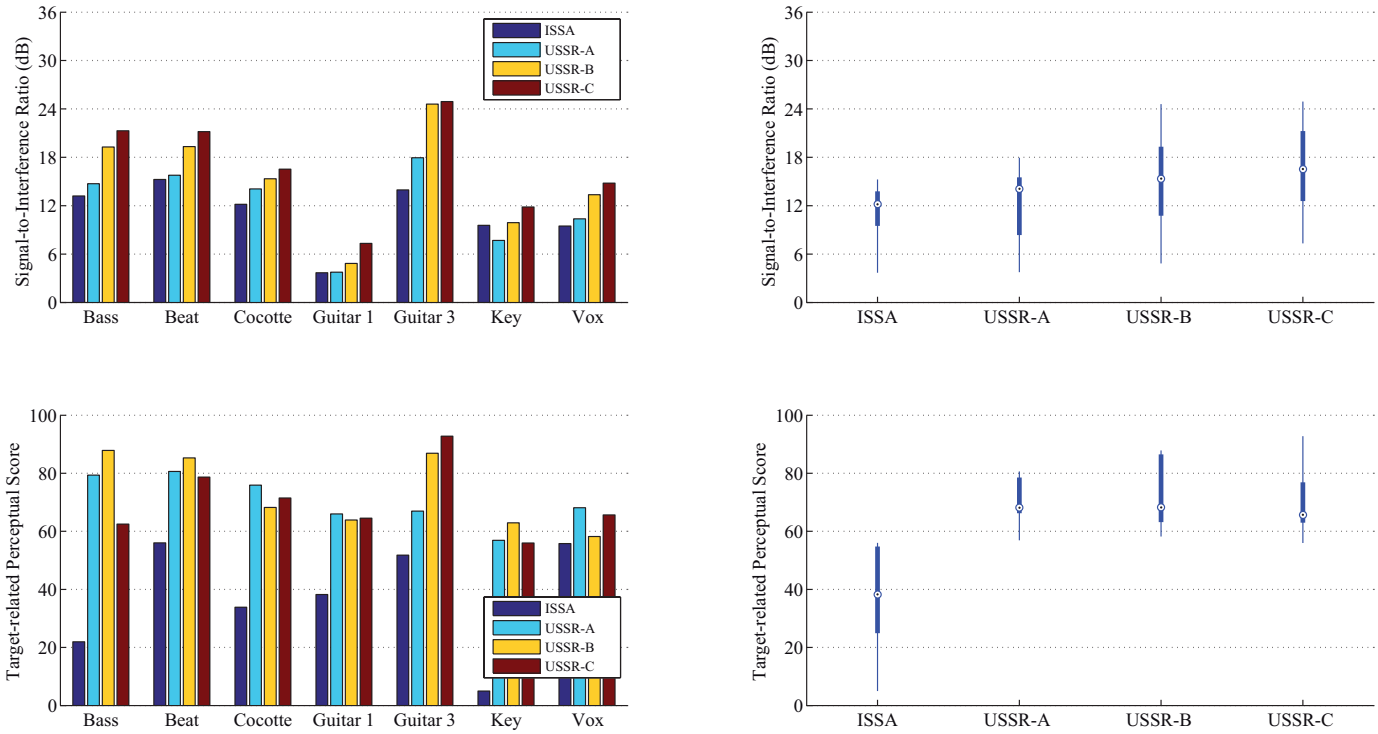

Fig. 5. Signal-to-interference ratios and target-related perceptual scores (left column) complemented by the median, the 25th and 75th percentiles, and outliers (right column) for an excerpt from "Lisztomania" by Phoenix.

hold true:

$$\sum_{l=1}^{d} \varrho_{il}^2 \sigma_l^2 < \sum_{l=1}^{d} \sigma_l^2, \qquad \sum_{v=u}^{d} \varrho_{uv}^2 \sigma_v^2 < \sum_{v=u}^{d} \sigma_v^2 \leqslant \sum_{v=1}^{d} \sigma_v^2. \tag{43}$$

The beamformer gain in (42) hence simplifies to

$$g_{io}^{\text{MMSE}} \overset{(43)}{\leqslant} \frac{\sigma_i^2 \sum_{l=1}^{d} \sigma_l^2}{\sum_{u=1}^{d} \sigma_u^2 \sum_{v=1}^{d} \sigma_v^2}$$

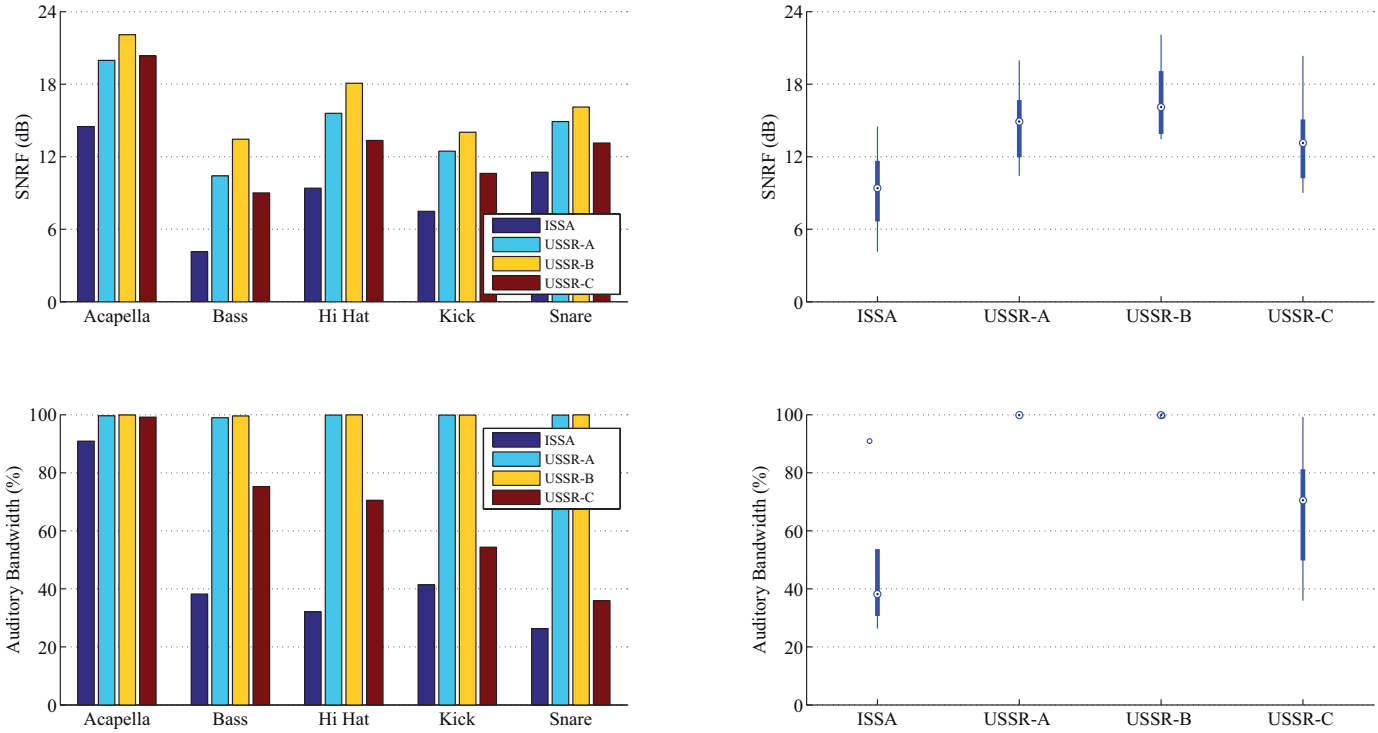

Fig. 6. Frequency-weighted signal-to-noise ratios and auditory bandwidths (left column) complemented by the median, the 25th and 75th percentiles, and outliers (right column) for an excerpt from "The Terrorist" by DJ Vadim.

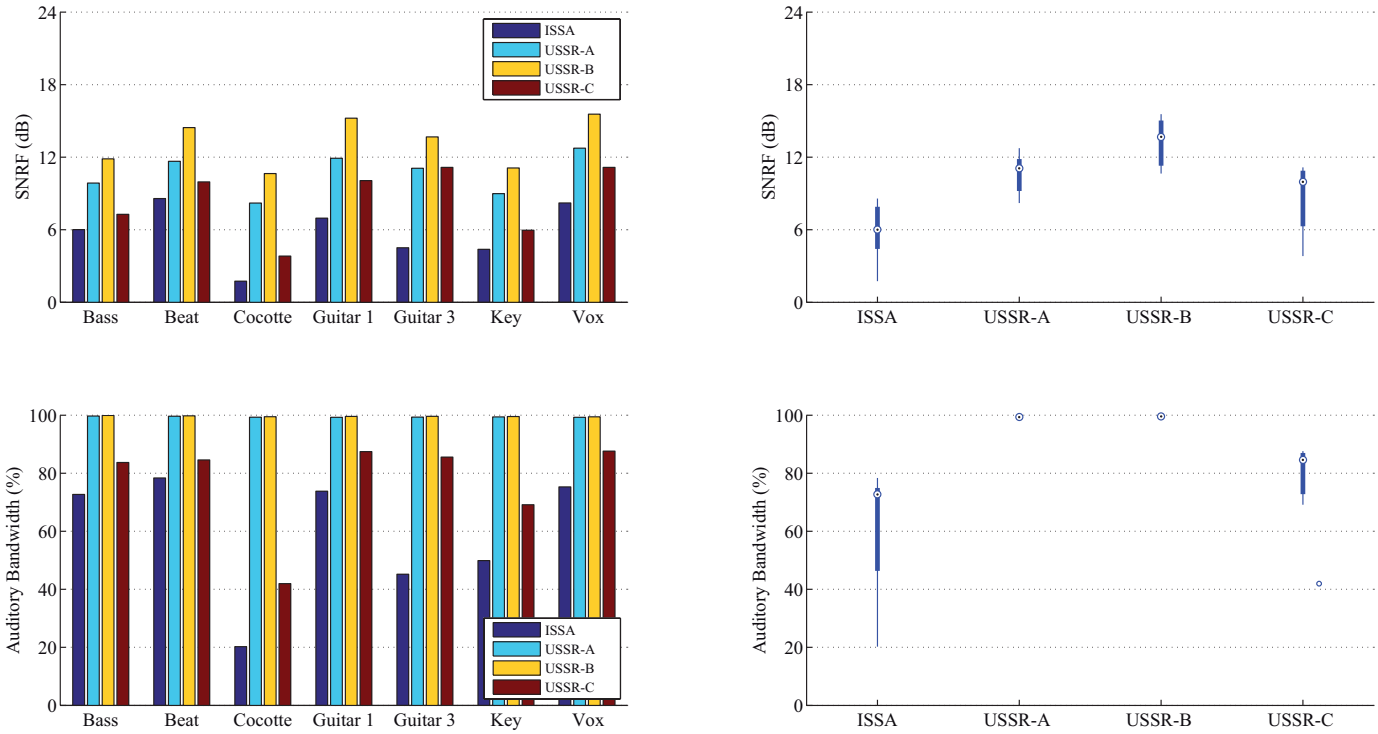

Fig. 7. Frequency-weighted signal-to-noise ratios and auditory bandwidths (left column) complemented by the median, the 25th and 75th percentiles, and outliers (right column) for an excerpt from "Lisztomania" by Phoenix.

$$\leqslant \frac{\sigma_i^2 \sum_{l=1}^{d} \sigma_l^2}{\left(\sum_{u=1}^{d} \sigma_u^2\right)\left(\sum_{v=1}^{d} \sigma_v^2\right)} \leqslant \frac{\sigma_i^2}{\sum_{u=1}^{d} \sigma_u^2}. \quad \blacksquare \tag{44}$$

Equation (44) underlines that just as the classical Wiener filter, the MMSE beamformer will attenuate the output signal at the

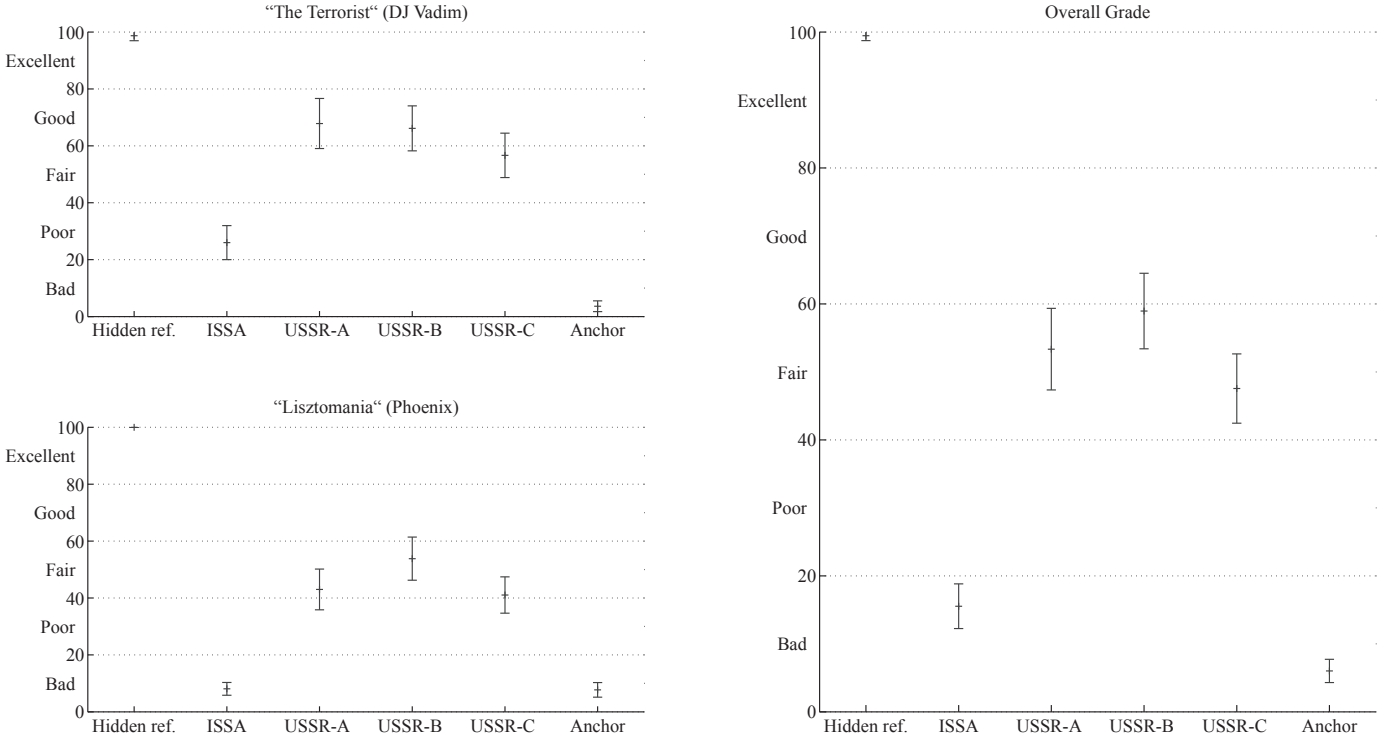

Fig. 8. Mean opinion scores and 95-% confidence intervals for the two music excerpts (left column) and the overall grades for the algorithms under test (right column).

attempt to minimize the error at TF points with a poor SIR. At worst, the attenuation may leave an audible spectral gap.

ACKNOWLEDGMENT

The authors would like to thank B. Dawson and DJ Vadim for providing the multitrack stems to the "The Terrorist" track from the "USSR: Life From The Other Side" LP as much as J. Pinel for the code for the data hiding technique.

REFERENCES

[1] 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, pp. 1721–1733, Aug. 2011.

[2] A. Liutkus *et al.*, "Informed source separation through spectrogram coding and data embedding," *Signal Process.*, vol. 92, pp. 1937–1949, Aug. 2012.

[3] 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.

[4] P. Comon and C. Jutten, *Handbook of Blind Source Separation: Independent Component Analysis and Applications*, 1st ed. Academic Press, 2010.

[5] S. Makino, T.-W. Lee, and H. Sawada, *Blind Speech Separation*. Springer, 2007.

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

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

[8] S. Rickard and O. Yılmaz, "On the approximate W-disjoint orthogonality of speech," in *Proc. IEEE ICASSP*, 2002, pp. 529–532.

[9] B. D. Van Veen and K. M. Buckley, "Beamforming: A versatile approach to spatial filtering," *IEEE Acoust., Speech, Signal Process. Mag.*, vol. 5, pp. 4–24, Apr. 1988.

[10] M. H. Er and A. Cantoni, "Derivative constraints for broad-band element space antenna array processors," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 31, pp. 1378–1393, Dec. 1983.

[11] S. Haykin, *Adaptive Filter Theory*, 4th ed. Prentice Hall, 2001.

[12] A. V. Oppenheim and R. W. Schafer, *Discrete-Time Signal Processing*, 3rd ed. Prentice Hall, 2010.

[13] H. Fastl and E. Zwicker, *Psychoacoustics: Facts and Models*, 3rd ed. Springer, 2007.

[14] B. R. Glasberg and B. C. J. Moore, "Derivation of auditory filter shapes from notched-noise data," *Hearing Res.*, vol. 47, pp. 103–138, Aug. 1990.

[15] V. G. Reju, S. N. Koh, and I. Y. Soon, "An algorithm for mixing matrix estimation in instantaneous blind source separation," *Signal Process.*, vol. 89, pp. 1762–1773, Sept. 2009.

[16] D. A. Huffman, "A method for the construction of minimum-redundancy codes," in *Proc. IRE*, 1952, pp. 1098–1102.

[17] J. M. Tribolet, P. Noll, B. J. McDermott, and R. E. Crochiere, "A study of complexity and quality of speech waveform coders," in *Proc. IEEE ICASSP*, 1978, pp. 586–590.

[18] V. Emiya, E. Vincent, N. Harlander, and V. Hohmann, "Subjective and objective quality assessment of audio source separation," *IEEE Trans. Audio Speech Lang. Process.*, vol. 19, pp. 2046–2057, Sept. 2011.

[19] ITU-R, *Method for the subjective assessment of intermediate quality level of coding systems (Rec. ITU-R BS.1534-1)*, Jan. 2003.

Stanislaw Gorlow (S'11) received the B.Eng. degree in information technology from Georg Simon Ohm University of Applied Sciences Nuremberg, Germany, in 2007 and the M.Sc. degree in electrical engineering and information technology from Ilmenau University of Technology, Germany, in 2010. He is currently pursuing a Ph.D. degree in the Computer Science Research Laboratory of Bordeaux (LaBRI) at University of Bordeaux 1, France.

He spent his military service as staff duty soldier with the German Army in Amberg between 2001–2002. During the period between 2005–2007 he went from being an intern to an employee at Dolby Germany, Nuremberg. In 2009, in parallel to his studies he also assisted in the Communications Research Laboratory (CRL) at the Institute for Information Technology in Ilmenau. His research interests are in digital speech and audio signal processing and related fields.

Mr. Gorlow is a member of the Audio Engineering Society (AES).

Sylvain Marchand (M'07—SM'11) received the M.Sc. degree and the Ph.D. degree both in computer science from University of Bordeaux 1, France, in 1996 and 2000, respectively.

From 2001 to 2011 he was with the Computer Science Research Laboratory of Bordeaux (LaBRI), University of Bordeaux 1, as an Associate Professor. During this period, he was a member of the Studio of Creation and Research in Computer Science and Electroacoustic Music (SCRIME). Since 2011, he is in charge of the "Image & Sound" curriculum at University of Western Brittany, France. His main research interests include sinusoidal modeling, spectrum analysis and synthesis, sound localization and spatialization, sound source separation and its application to active listening.

Prof. Marchand is a member of the scientific committee of the international conference on Digital Audio Effects (DAFx) and he is the head of the DReaM project (see http://dream.labri.fr). He was furthermore an Associate Editor of the IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING between 2007–2011.