INFORMED SOURCE SEPARATION FOR STEREO UNMIXING — AN OPEN SOURCE IMPLEMENTATION

Sylvain Marchand
L3i
University of La Rochelle
France
sylvain.marchand@univ-lr.fr

Pierre Mahé
LIS
University of Toulon
France
pierre.mahe@univ-tln.fr

ABSTRACT

Active listening consists in interacting with the music playing and has numerous potential applications from pedagogy to gaming, through creation. In the context of music industry, using existing musical recordings (e.g. studio stems), it could be possible for the listener to generate new versions of a given musical piece (i.e. artistic mix). But imagine one could do this from the original mix itself. In a previous research project, we proposed a coder / decoder scheme for what we called informed source separation: The coder determines the information necessary to recover the tracks and embeds it inaudibly (using watermarking) in the mix. The decoder enhances the source separation with this information. We proposed and patented several methods, using various types of embedded information and separation techniques, hoping that the music industry was ready to give the listener this freedom of active listening. Fortunately, there are numerous other applications possible, such as the manipulation of musical archives, for example in the context of ethnomusicology. But the patents remain for many years, which is problematic. In this article, we present an open-source implementation of a patent-free algorithm to address the mixing and unmixing audio problem for any type of music.

1. INTRODUCTION

Active listening of music is an artistic as well as a technological topic of growing interest, that concerns offering listeners the possibility to interact in real time with the music, e.g. to modify the elements, the sound characteristics, and the structure of the music while it is played. This involves, among other examples, advanced remixing processes such as generalized karaoke (muting any musical element, not only the lead vocal track), respatialization, or upmixing. The applications are numerous, from learning / teaching of music to gaming, through new creative processes (disc jockeys, live performers, etc.). In the context of ethnomusicological archiving, the recordings can consist of several tracks, but for the purpose of compatibility, only the mix can often be distributed in the archive. Thus, a technique allowing the user to get access back to the separate tracks from the stereo mix can be very useful.

To get this new freedom, a simple solution would be to give the user access to the individual tracks that compose the mix, by storing them into some multi-track format. This approach has two main drawbacks: First, it leads to larger multi-track files. Second, it yields files that are not compatible with the prevailing stereo standards. Another solution is to perform some blind separation of the sources from the stereo mix. The problem is that even with state-of-the-art blind source separation techniques the quality is usually insufficient and the computation is heavy (see [1]).

In the DReaM project (see [2]), we proposed an Informed Source Separation (ISS) approach (see [3]) to accurately recover the separate tracks from the stereo mix. The present article will focus on this approach only, where the system consists of a coder and a decoder. The coder is used at the mixing stage, where the separate tracks are known. It determines the information necessary to recover the tracks from the mix and embeds it in the mix. In the classic case of Pulse-Code Modulation (PCM), this information is inaudibly hidden in the mix by a watermarking technique. With a legacy system, the coded stereo mix can be played and sounds just like the original, although it includes some additional information. Apart from backward compatibility with legacy systems, a further advantage concerns the fact that the file size stays comparable to the one of the original mix. The decoder performs source separation of the mix with parameters given by the additional information. This ISS approach permits producing good separate tracks, thus enabling active listening applications.

The original target of the DReaM project was the music industry, which turned out to be quite conservative. For instance, the actors of music industry appeared to be reserved with the use of audio formats that are alternative to conventional stereo encoding, hence hindering the development of object-based formats or advanced spatial audio formats such as Ambisonics. Another example is the fact that listeners are considered as (passive) consumers, even if some want to behave as musicians (active listeners, content producers, etc.).

Fortunately, there is some opportunity for the system developed in the project for an application to musical archives (see [4]). Indeed, some recordings contain several tracks, but the diffusion format is still legacy stereo. Thus, having a format backward compatible with standard stereo but allowing to recover the individual tracks present in the mix can be of interest. The DReaM project showed that it is possible. However, since the finality of the project was industrial, the ISS methods were patented. For new – non commercial – applications, a patent-free method was needed. The contribution of the present article is the definition and implementation of such a method.

The remainder of this article is organized as follows. Section 2 presents the DReaM project: its fundamentals and target applications. Section 3 describes the separation / unmixing methods developed in the project, whereas Section 4 introduces a patent-free method: ReaLiTy. Finally, Section 5 draws some conclusions.

The DReaM project was supported by the French ANR *(Agence Nationale de la Recherche)*, from 2009 to 2014 (ANR-09-CORD-006).

2. THE DREAM PROJECT

DReaM is a French acronym for *"le Disque Repensé pour l'écoute active de la Musique"*, which means "the disc thought over for active listening of music". This is the name of an academic project (2009–2014) with industrial finality, coordinated by the first author, and funded by the French National Research Agency (ANR). The project involved academic partners (LaBRI – University of Bordeaux, Lab-STICC – University of Brest, GIPSA-Lab – Grenoble INP, LTCI – Telecom ParisTech, ESPCI – Institut Langevin) together with iKlax Media, a company for interactive music that contributed to the Interactive Music Application Format (IMAF) standard (see [5]).

The origin of the project comes from the observation of artistic practices. More precisely, composers of acousmatic music conduct different stages through the composition process, from sound recording (usually stereophonic) to diffusion (multiphonic). During live interpretation, they interfere decisively on spatialization and coloration of pre-recorded sonorities. For this purpose, the musicians generally use a mixing console to upmix the musical piece being played from an audio CD. This requires some skills, and imposes musical constraints on the piece. Ideally, the individual tracks should remain separate. However, this multi-track approach is hardly feasible with a typical (stereophonic) audio CD.

Nowadays, the audience is more eager to interact with the musical sound. Indeed, more and more commercial CDs come with several versions of the same musical piece. Some are instrumental versions (e.g. for karaoke), other are remixes. The karaoke phenomenon gets generalized from voice to instruments, in musical video games such as *Rock Band*. But in this case, enabling interaction translates to users having to buy a video game, which includes the multi-track recording.

Yet, the music industry seems to be reluctant to releasing the multi-track versions of big-selling hits. The only thing the user can get is a standard CD, thus a stereo mix, or its digital version available for download or streaming, now that the physical version (at least the CD) disappears.

2.1. Objectives

In general, the project aims at solving a so-called inverse problem, to some quality extent, at the expense of additional information. In particular, an example of such an inverse problem can be source separation: recovering the individual source tracks from the given mix.

On the one hand coding the solution (e.g. the individual tracks and the way to combine them) can bring high quality, but with a potentially large file size, and a format not compatible with existing stereo formats. On the other hand the blind approach (without information) can produce some results, but of insufficient quality for demanding applications (see [1]). The blind approach can be regarded as an estimation without information, while coding can be regarded as using information (from each source) without any estimation (from the mix).

The informed approach proposed by DReaM is just in between these two extremes: getting musically acceptable results with a reasonable amount of additional information. The problem is now to identify and encode efficiently this additional information. Remarkably, ISS can thus be seen both as a multi-track audio coding scheme using source separation, or as a source separation system helped by audio coding.

This approach addresses the source separation problem in a coder / decoder configuration. At the coder (see Figure 1), the additional information is estimated from the original source signals before the mixing process and is inaudibly embedded into the final mix. At the decoder (see Figure 2), this information is extracted from the mix and used to assist the separation process.

So, a solution can be found to any problem, thanks to the additional information embedded in the mix.

> "There's not a problem that I can't fix,
> 'cause I can do it in the mix!"
> (Indeep – Last Night a DJ Saved my Life)

The original goal of the project was to propose a fully backward-compatible audio-CD permitting musical interaction.

The idea was to inaudibly embed (using a high-capacity watermarking technique) in the audio track some information enabling to some extent the musical decomposition, that is the inversion of the music production chain: dynamics decompression, source separation (unmixing), deconvolution, etc.

With a standard CD player, one would listen to the fixed mix. With an active player however, one could modify the elements and the structure of the audio signal while listening to the music piece.

Now that the music is getting all digital, the consumer gets access to audio files instead of physical media. In this article we will consider only audio files without compression.

2.2. Applications

Active listening (see [6]) amounts to performing various operations that modify the elements and structure of the music signal during the playback of a piece. This process, often simplistically called remixing, includes generalized karaoke, respatialization, or applying certain effects to individual audio tracks (e.g. adding some distortion to an acoustic guitar). The goal is to enable the listener to enjoy freedom and personalizing of the musical piece through various reorchestration techniques. Alternatively, active listening solutions intrinsically provide simple frameworks to the artists to produce different versions of a given piece of music. Moreover, it is an interesting framework for music learning / teaching applications.

2.2.1. Respatialization

The original application was to let the public experience the freedom of composers of electroacoustic music during their live performances: moving the sound sources in the acoustic space. Although changing the acoustical scene by means of respatialization is a classic feature of contemporary art (electroacoustic music), and efforts have been made in computer music to bring this practice to a broader audience (see [7]), the public seems just unaware of this possibility and rather considered as passive consumers by the music industry. However, during the public demonstrations of the DReaM project, we felt that the public was very reactive to this new way of interacting with music, to personalize it, and was ready to adopt active listening, mostly through musical games.

2.2.2. Generalized Karaoke

The generalized karaoke application is the ability to suppress any audio source, either the voice (classic karaoke) or any instrument ("music minus one"). The user can then practice singing or playing

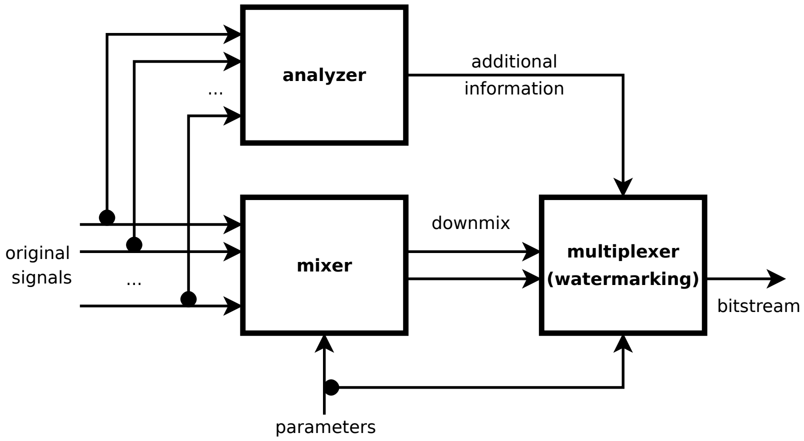


Figure 1: *Architecture of an Informed Source Separation (ISS) coder.*

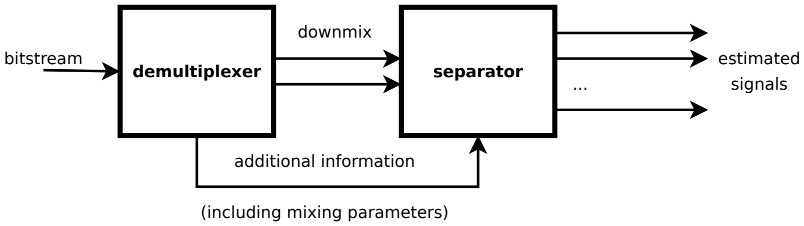


Figure 2: *Architecture of an Informed Source Separation (ISS) decoder.*

an instrument while being integrated in the original mix and not a cover song.

Note that these two applications (respatialization and generalized karaoke) are related, since moving a source far away from the listener will result in its muting, and reciprocally the ability to mute sources can lead to the monophonic case (the spatial image of a single source isolated) where respatialization is much easier (possible to some extent even without recovering the audio object from this spatial image).

2.2.3. Sound Archives

It turns out that the system developed in the project might be very useful for musical archives. Indeed, some recordings contain several tracks, but the diffusion format is still legacy stereo. Thus, having a format backward compatible with standard stereo but allowing to recover the individual tracks present in the mix can be of interest.

3. INFORMED SOURCE SEPARATION METHODS

A stereo (2-channel) mixture $\{y_c(n)\}_{c=1,2}$ will be produced from K source signals $\{x_k(n)\}_{k=1}^{K}$ and panning angles θ_k (which are not azimuths, see [8] for details), the latter leading to a mixing matrix A where A_{ck} denotes the contribution of the kth input source to the cth output channel. In this article, we will consider a simple case where the mixing matrix A is obtained from panning angles θ using Equations (1) and (2)

$$A_{1k} = \sin(\theta_k) \quad (1)$$

$$A_{2k} = \cos(\theta_k) \quad (2)$$

such that $A_{1k}^2 + A_{2k}^2 = 1$ (energy conservation). The values for the panning angles θ will range from 0 (right) to $\pi/2$ radians (left).

Source separation then consists in recovering (estimates of) the source signals x_k from the mix signals y_c, possibly with the help of additional information extracted from x_k (informed approach).

Over the years of the project, several Informed Source Separation (ISS) methods were proposed. More precisely, this section presents the similarities, differences, strengths, and weaknesses of four of them. A detailed technical description or comparison is out of the scope of this article. Instead, we will propose a new – free – method, which is a mix of the original methods. The detailed descriptions of the following methods can rather be found in [9], [10], [8], and [11], while their comparison is done in [12].

The majority of the ISS methods aims at extracting the contribution of each source from each Time-Frequency (TF) point of the mix, at least in terms of magnitude, and sometimes phase too.

3.1. Local Inversion

The first method performs a local inversion (see [9] and [13]) of the mix for each TF point, using the information of the two predominant sources in this point (as well as the knowledge of the mixing matrix). More precisely, at each TF point two sources can be reconstructed from the two (stereo) channels, by a local two-by-two inversion of the mixing matrix. This way, we get estimates of the magnitude and phase of the prominent sources. But the problem is that the remaining $K-2$ sources exhibit a spectral hole (no estimated signal), which is perceived as quite annoying in subjective listening tests (see [8]). Also, this method requires the mixing matrix A to be of rank K.

3.2. Minimum Mean-Square Error Filtering

The second method performs classic Minimum Mean-Square Error (MMSE) filtering (see [10] and [14]) using Wiener filters driven by the information about the power of the sources (as well as the mixing matrix), the corresponding spectrograms being transmitted using either sound (NMF) or image (JPG) compression techniques (see [12] for details). In contrast to the local inversion method, MMSE does not constrain as much the mixing matrix A and is therefore more flexible towards the mixing configurations. The separation quality, however, is much better when A is of rank K.

3.3. Linearly Constrained Spatial Filtering

The third method is called Undetermined Source Signal Recovery (USSR), and performs linearly constrained spatial filtering (see [8] and [15]) using a Power-Constraining Minimum-Variance - (PCMV) beamformer, also driven by the information about the power of the sources (and their spatial distribution) and ensuring that the output of the beamformer matches the power of the sources (additional information transmitted in ERB/dB scales, see Section 4.1.2). In the stereo case, if only two predominant sources are detected, the beamformer is steered such that one signal component is preserved while the other is canceled out. Applying this principle for both signal components results in inverting the mixing matrix (first method). Moreover, dropping the power constraint will turn the PCMV beamformer into an MMSE beamformer (second method). Otherwise, the PCMV beamformer takes advantage of the spatial distribution of the sources to produce better estimates.

3.4. Iterative Phase Reconstruction

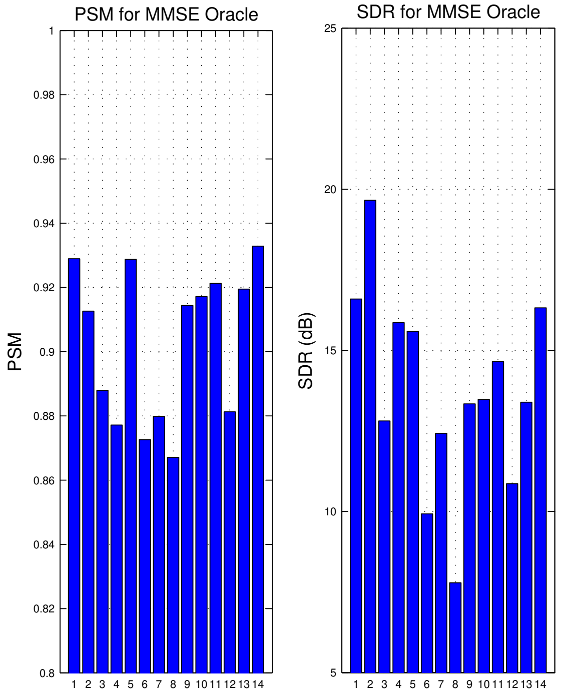
The fourth method performs iterative phase reconstruction and is called IRISS (Iterative Reconstruction for Informed Source Separation), see [11]. It also uses the magnitude of the sources (transmitted in ERB/dB scales) as well as a binary activity map as an additional information to the mix. The main point of the method is to constrain the iterative reconstruction of all the sources so that Equation (5) is satisfied at each iteration very much like the Multiple Input Spectrogram Inversion (MISI) method (see [16]). Contrary to MISI, both amplitude and phase of the STFT are reconstructed in IRISS, therefore the remix error should be carefully distributed. In order to do such a distribution, an activity mask derived from the Wiener filters is used. The sources are reconstructed at the decoder with an initialization conditioned at the coding stage. It is noticeable that this technique was specifically designed for mono mixtures (1-channel), where it gives the best results.

3.5. Evaluation

3.5.1. Performances

The quality performance of the system reaches the needs of many real-life applications (for each of the four methods described above). The comparison of the four original methods can be found in [12], for the linear instantaneous and convolutive case, using either the objective Signal-to-Distortion Ratio (SDR) criterion of BSSEval (see [17]) or the subjective Perceptual Similarity Measure (PSM) of PEMO-Q (see [18]), closer to perception. A set of 14 musical excerpts from the Quaero database has been considered (see [12] for details).

Figure 3 shows the performances of MMSE (Wiener) filtering with access to full information (oracle situation) about sound sources for both subjective (PSM) and objective (SDR) measures. The PSM is often above 0.9 (1 corresponding to perfection), and the SDR is around 15dB (which is quite good).



(a) Subjective results (PSM) (b) Objective results (SDR)

Figure 3: *Performances of MMSE filtering with access to full information (oracle) about sound sources (estimated signals compared to the original signals).*

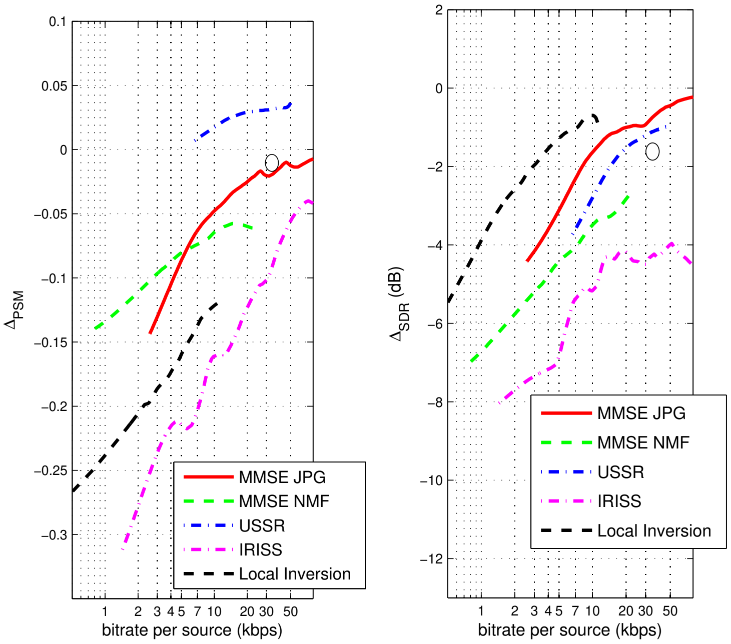
Figure 4 shows the relative performances of the DReaM methods, relatively to the MMSE oracle, as functions of the additional information bitrate. It turns out that the first method (local inversion) exhibits the best objective (SDR) results, while the third method (USSR) exhibits the best subjective (PSM) results; this was also verified in a formal listening test conducted in [8].

It is important to note that the complexity of these methods is low, allowing real time. Moreover, as shown in [12], the typical bitrates for the additional information are approximately 5 to 10 kbits per second for each source, which is quite reasonable.

The problem with these methods is that they are protected with patents.

3.5.2. Patents

The patent of the first method (see [13]) protects the local inversion technique as well as the encoding of the active sources indices. The patent of the second method (see [14]) protects the coding of the additional information, but not Wiener filtering which is a well-known technique. The patent of the third method (see [15]) protects the use of the PCMV beamformer for source separation, whereas the ERB and dB scales used for the additional information reduction are well-known, and also used by the fourth method.



(a) Subjective results (PSM)

(b) Objective results (SDR)

Figure 4: *Performances of the DReaM methods (relatively to the MMSE oracle), as functions of the additional information bitrate. The black circle indicates the performance of the proposed ReaLiTy implementation.*

This last method is patent-free, but unfortunately not suitable for stereo mix, with a lower quality and a higher complexity (increased processing time).

Thus, there is room for an efficient patent-free method, with additional information represented in ERB/dB scales and filtering performed using the standard Wiener (MMSE) filtering, provided the watermarking technique used is also patent-free.

This method, called ReaLiTy, is described in the next section. The code is distributed as free software, and comes with a sound example. The performance of the proposed method on that specific example is indicated[1] by a black circle on Figure 4. The subjective result is close to MMSE oracle performance thanks to the high bitrate per source (35 kbps) for the additional information, but stays below this limit unlike USSR (whose filtering is not MMSE). The objective result is comparable to those of USSR and MMSE JPG (the bitrate being unfortunately not suitable for MMSE NMF), although slightly below.

4. REALITY: A FREE IMPLEMENTATION

All the methods developed during the DReaM project are based on a coder / decoder scheme. The coder produces a stereo mix from the K source signals using panning angles, and the decoder recovers (estimates of) these signals from the mix, using the additional information inaudibly embedded by the coder. This section gives the details for the coder and the decoder of the proposed ReaLiTy method, which is patent-free and comes with a free software implementation in Python programming language.

The source signals $\{x_k(n)\}_{k=1}^{K}$ are block-wise time-frequency mapped by means of the Short-Time Fourier Transform (STFT) using Equation (3)

$$X_k(h,b) = \sum_{n=0}^{N-1} x_k(hH+n)w(n)e^{-\jmath 2\pi nb/N} \qquad (3)$$

where $0 \leq b < N$ is the frequency index, N is the frame size, h is the frame index, H is the hop size, and $\jmath$ is the imaginary unit. In practice, for w we use the Hann window of size $N = 2048$, for a sampling frequency $F_s = 44100$Hz. We will allow a 50% overlap, thus $H = N/2$.

4.1. Coder

As shown on Figure 1, the coder consists of three building blocks: mixer, analyzer, and multiplexer.

4.1.1. Mixer

The sources are defined as K mono signals x_k of same length L, with sampling rate F_s. The mixer takes these original signals x_k and panning angles θ_k and produces a stereo (2-channel) mixture $\{y_c(n)\}_{c=1,2}$.

[1] However this is only an indication, since this figure was originally generated for [12] with different sound excerpts, and unfortunately more than 10 years after it was impossible to get access to them. More precisely, if the database is known, the excerpts were not specified and this information is apparently lost now. Running a new comparison on different excerpts turned out to be also impossible, since we do not have access to the (patented) code of the original methods anymore.

We first consider linear and time-invariant mixing systems. Formally, we suppose that each source signal x_k is mixed into each destination channel c through the use of some mixing coefficients a_{ck}, leading to Equation (4).

$$y_c(n) = \sum_{k=1}^{K} a_{ck} \cdot x_k(n) \tag{4}$$

Since the mixing coefficients are constant over time the mixing is said to be linear instantaneous.

If the mixing coefficients a_{ck} are replaced by filters, and the product in Equation (4) is replaced by the convolution, the mixing is then said to be convolutive. We can easily handle this case (see [12]) with the STFT representation if the length of the mixing filters is sufficiently short compared to the window size, thanks to the convolution theorem, with Equation (5)

$$Y_c(h, b) \approx \sum_{k=1}^{K} A_{ck}(b) \cdot X_k(h, b) \tag{5}$$

where $A_{ck}(b)$ is understood as the frequency response of filter a_{ck} at frequency index b. When the mixing process is linear instantaneous and time invariant, A_{ck} is constant and the $2\times K$ matrix A is called the mixing matrix. The mixing process can thus be rewritten as matrix multiplication in Equation (6)

$$Y \approx A \cdot X \tag{6}$$

where $Y = [Y_1, Y_2]^\top$ and $X = [X_1, \cdots, X_K]^\top$ are column vectors respectively gathering all mixtures and sources at the time-frequency (TF) point (h, b).

4.1.2. Analyzer

The analyzer also takes the original signals x_k as inputs, to compute the additional information to be embedded in the mix.

At the origin of the DReaM project, this information consisted of the indices of the two most prominent sources, that is the two sources with the highest energy at the considered TF point, since this information can be used to solve the interference of the sources at this point, by local inversion (see [9]). This information can be efficiently coded with $\lceil \log(K(K-1)/2) \rceil$ bits per TF point. But the local inversion technique is patented (see [13]).

The information about the power spectrum of each source turned out to be extremely useful and more general. Indeed, if we know the power of all the sources, we can determine the two predominant sources. We can also derive activity patterns for all the sources. As shown in [10], this information can be coded using sound or image compression techniques. The problem, again, is that it is patented (see [14]).

Let us consider the instantaneous Power Spectral Density (PSD) $E_k(h, b)$, calculated according to Equation (7).

$$E_k(h, b) = |X_k(h, b)|^2 \tag{7}$$

Fortunately, this information can efficiently be coded on a double-logarithmic scale using simple psychoacoustic considerations. More precisely, a significant reduction of this information can be achieved in two ways: first, by reducing the frequency resolution of the PSDs $E_k(h, b)$ in approximation of the critical bands (see [19]), and second, by quantizing the obtained PSD values $\hat{E}_k(h, z)$ with a step size equal to some value Δ, which is put in relation to an appropriate psychoacoustic criterion.

Scaling. 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 of [20] we can define a relation between the frequency index b and the critical-band index z_b by Equation (8)

$$z_b = \lfloor 21.4 \log_{10}(4.37 b(F_s/1000)/N + 1) \rfloor \tag{8}$$

where $\lfloor \cdot \rfloor$ is the floor function. The zth critical-band value of the approximate PSDs is then calculated as the arithmetic mean between $\text{lower}(z) = \inf\{b \colon z_b = z\}$ and $\text{upper}(z) = \sup\{b \colon z_b = z\}$ according to Equation (9).

$$\bar{E}_k(h, z) = \frac{1}{\text{upper}(z) - \text{lower}(z) + 1} \sum_{b=\text{lower}(z)}^{\text{upper}(z)} E_k(h, b) \tag{9}$$

Recovering the Short-Time PSDs (STPSDs) in linear scale (to the resolution of the STFT) is as easy as Equation (10).

$$E_k(h, b) \approx \bar{E}_k(h, z_b) \tag{10}$$

Quantization. Furthermore, under the assumption that the the minimum just-noticeable-difference level and so the maximum allowed quantization error is 1dB (see [19]), the quantization step size Δ is chosen as 2dB, and the irrelevancy-reduced PSD values are obtained from the uniform quantizer in Equation (11)

$$\bar{E}_k^{\Delta}(h, z) = [5 \log_{10} \bar{E}_k(h, z)] \tag{11}$$

where $[\cdot]$ denotes the round-to-nearest rounding function. Note that replacing 5 by 10 in the previous equation would lead to the classic dB scale. Recovering the STPSD values in linar scale (by "dequantization") is as easy as Equation (12).

$$\bar{E}_k(h, z) \approx 10^{\bar{E}_k^{\Delta}(h,z)/5} \tag{12}$$

4.1.3. Multiplexer

The multiplexer takes the downmix y_c as well as the mixing parameters (panning angles) and the additional information as inputs, in order to produce a bitstream: the resulting stereo sound file.

The panning angles θ_k are simply rounded to the nearest integer value and quantized on 8 bits. The additional information consists of the STPDSs of the K source signals, $\bar{E}_k^{\Delta}$, quantized on B_S bits. Increasing B_S will lead to a better audio quality, but at the expense of a greater number of bits necessary at each STFT frame to encode the full additional information (including mixing parameters): $(8 + W \times B_S) \times K$ bits, where W is the number of bands of the ERB scale. In our implementation, we use $W = 136$, $B_S = 6$, for $K = 5$ sources, for a total of 4120 bits per frame.

These bits will be embedded inaudibly using watermarking. To avoid any patent, we will consider the most basic technique consisting in hiding the data in the Least Significant Bits (LSBs) of the downmix samples.

Since this downmix is stereo (2 channels), and each STFT frame consists of N samples, with a hop size of $H = N/2$ meaning 50% overlap, if we use B_C LSB bits per channel we can hide $B_C \times N$ bits per frame. In our implementation we use $N = 2048$ and $B_c = 3$, for a total of 6144 bits available.

With these settings, we could handle up to 7 sources. To go further, one can increase B_C or decrease B_S. In the first case, the quality of the mix will begin to degrade (the watermark becoming audible), and the second case the quality of the estimated source signals will degrade. Using a higher capacity watermarking or entropy coding for the data could also be solutions.

4.2. Decoder

As shown on Figure 2, the decoder consists of two building blocks: demultiplexer and separator.

4.2.1. Demultiplexer

From the input bitstream, the demultiplexer has to recover the downmix plus the additional information (including mixing parameters). As an approximation, the downmix y_c will be the stereo signal of the input. The additional information (θ_k and $\bar{E}_k^{\Delta}$) is simply extracted from the LSBs of the samples of this input signal.

4.2.2. Separator

The core block of the decoder is the separator, aiming at estimating the K original signals x_k from the downmix y_c and this additional information, consisting of the K source STPSDs in ERB/dB scale ($\bar{E}_k^{\Delta}$) together with the mixing parameters (θ_k), leading to the mixing matrix A (see Section 3).

Filtering. As shown in Section 3.5, if one wants to maximize the objective quality (SDR), one could use the first DReaM method (local inversion) but then mess with a patent ([13]), and if one wants to maximize the subjective quality (PSM), one could use the third method (USSR) but then mess with another patent (see [15]). The Wiener (MMSE) filtering used by the second method is a good compromise, and patent-free. This filtering is done according to Equation (13).

$$\hat{X}_k(h,b) = \sum_{c=1}^{2} Y_c(h,b) \cdot \frac{A_{ck} \cdot E_k(h,b)}{\sum_{s=1}^{K} A_{cs} \cdot E_s(h,b)} \tag{13}$$

Adjusting. Since the STPSDs of the sources E_k are known, we can scale the estimated source spectra $\hat{X}_k$ to adjust these STPSDs. The spectra $\left\{\hat{X}_k(h,b)\right\}_{k=1}^{K}$ are then transformed back to the time domain to get the signals $\{\hat{x}_k(n)\}_{k=1}^{K}$ using the inverse STFT (ISTFT) with a classic overlap-add (OLA) procedure, with 50% overlap ($H = N/2$), the Hann window used for the STFT ensuring perfect reconstruction of the signals in this case.

In practice, it could be a good idea (in case of non-linear spectral processing) to apply the window w at both STFT and ISTFT stages, using the square root of the Hann window (so that the product of the windows of the two stages results in the original Hann window). This is done is our free software implementation[2], programmed in Python.

5. CONCLUSION

Originally thought as a way to interact with the music signal through its real-time decomposition / manipulation / recomposition, in the DReaM project the emphasis has been laid on the mixing stage, leading to source separation / unmixing techniques using additional information to improve the quality of the results. DReaM can also be regarded as a multi-track coding system based on source separation.

The initial aim was to give freedom to the listener, in the context of music industry, but artistic as well as industrial problems arose. For example, the artwork is sacred – it shall not be “altered”. Also, the method requires studios recordings – involving copyright issues with studios / producers / majors. Finally, the method requires mastering the whole production chain – meaning entering Digital Audio Workstations (DAWs), which can hardly be done. But DReaM has shown the possibility to produce a mix allowing source separation, backward compatible with legacy stereo, thus without the need of some multi-track format. Unfortunately, the industrial finality of the project led to patents on the original methods.

In this article, we proposed ReaLiTy, a patent-free version of the system. It is based on well-known techniques such as LSB watermarking, ERB/dB scale, or Wiener filtering. A free software implementation in Python programming language is available online. We hope that it could be used e.g. for storing / spreading multi-track sound archives within the standard stereo format, or could serve as a basis for future research.

6. ACKNOWLEDGMENTS

“You may say I’m a dreamer, but am not the only one.” (John Lennon – Imagine). Thus, the first author would like to thank all the members of the project consortium for having made the DReaM come true. Thanks must also go to Jim Beauchamp, for his interest in music signal processing and his invitations to meetings of the Acoustical Society of America (ASA), and to Rolf Bader, who saw there the opportunity of the DReaM methods for academic applications, and allowed the first version of ReaLiTy to be developed in Hamburg, Germany. Stanislaw Gorlow provided the Matlab implementation of the USSR method, which served as a basis for this first Python development by Pierre Mahé, which in turn served for the final version of ReaLiTy by Sylvain Marchand.

[2] ReaLiTy: https://www.sylvain-marchand.info/ReaLiTy/

7. REFERENCES

[1] Pierre Comon and Christian Jutten, Eds., *Handbook of Blind Source Separation – Independent Component Analysis and Applications*, Academic Press, 2010.

[2] Sylvain Marchand, Roland Badeau, Cléo Baras, Laurent Daudet, Dominique Fourer, Laurent Girin, Stanislaw Gorlow, Antoine Liutkus, Jonathan Pinel, Gaël Richard, Nicolas Sturmel, and Shuhua Zhang, "DReaM: A novel system for joint source separation and multi-track coding," in *Proceedings of the 133rd AES Convention*, San Francisco, California, USA, October 2012.

[3] Kevin H. Knuth, "Informed source separation: A Bayesian tutorial," in *Proceedings of the European Signal Processing Conference (EUSIPCO)*, Antalya, Turkey, September 2005.

[4] Sylvain Marchand, "Spatial manipulation of musical sound: Informed source separation and respatialization," in *Computational Phonogram Archiving (Current Research in Systematic Musicology)*, Rolf Bader, Ed., pp. 175–190. Springer Nature, Switzerland, 2019.

[5] ISO/IEC, *ISO/IEC 23000-12, Information technology – Multimedia application format (MPEG-A) – Part 12: Interactive Music Application Format (IMAF)*, 2010.

[6] Philippe Lepain, "Écoute interactive des documents musicaux numériques," in *Recherche et applications en informatique musicale*, Marc Chemillier and François Pachet, Eds., pp. 209–226. Hermes, Paris, France, 1998, In French.

[7] François Pachet and Olivier Delerue, "A constraint-based temporal music spatializer," in *Proceedings of the ACM Multimedia Conference*, Brighton, United Kingdom, 1998.

[8] Stanislaw Gorlow and Sylvain Marchand, "Informed audio source separation using linearly constrained spatial filters," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 1, pp. 3–13, January 2013.

[9] Mathieu Parvaix and Laurent Girin, "Informed source separation of linear instantaneous under-determined audio mixtures by source index embedding," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 6, pp. 1721–1733, August 2011.

[10] Antoine Liutkus, Jonathan Pinel, Roland Badeau, Laurent Girin, and Gaël Richard, "Informed source separation through spectrogram coding and data embedding," *Signal Processing*, vol. 92, no. 8, pp. 1937–1949, August 2012.

[11] Nicolas Sturmel and Laurent Daudet, "Informed source separation using iterative reconstruction," *IEEE Transactions on Audio, Speech, and Language Processing*, , no. 1, pp. 178–185, January 2013.

[12] Antoine Liutkus, Stanislaw Gorlow, Nicolas Sturmel, Shuhua Zhang, Laurent Girin, Roland Badeau, Laurent Daudet, Sylvain Marchand, and Gaël Richard, "Informed audio source separation: A comparative study," in *Proceedings of the European Signal Processing Conference (EUSIPCO)*, Bucharest, Romania, August 2012.

[13] Mathieu Parvaix, Laurent Girin, Jean-Marc Brossier, and Sylvain Marchand, *Method and Device for Forming a Mixed Signal, Method and Device for Separating Signals, and Corresponding Signal*, FR2944403, EP2417597, US20120203362, KR1020120006050, JP2012523579, WO2010116068 (patents), France, Europe, United States, Korea, Japan, World, 2010.

[14] Laurent Girin, Antoine Liutkus, Gaël Richard, and Roland Badeau, *Method and Device for Forming a Digital Audio Mixed Signal, Method and Device for Separating Signals, and Corresponding Signal*, FR2966277, EP2628154, US20140037110, WO2012049176 (patents), France, Europe, United States, World, 2012.

[15] Sylvain Marchand and Stanislaw Gorlow, *Method and Device for Separating Signals by Minimum Variance Spatial Filtering Under Linear Constraint*, FR2996043, EP2901447, US20150243290, JP2015530619, WO2014048970 (patents), France, Europe, United States, Japan, World, 2014.

[16] David Gunawan and Deep Sen, "Iterative phase estimation for the synthesis of separated sources from single-channel mixtures," *IEEE Signal Processing Letters*, vol. 17, no. 5, pp. 421–424, May 2010.

[17] Emmanuel Vincent, Rémy Gribonval, and Cédric Févotte, "Performance measurement in blind audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1462–1469, July 2006.

[18] Rainer Huber and Birger Kollmeier, "PEMO-Q – a new method for objective audio quality assessment using a model of auditory perception," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 6, pp. 1902–1911, November 2006.

[19] Hugo Fastl and Eberhard Zwicker, *Psychoacoustics: Facts and Models*, Springer, third edition, 2007.

[20] Brian R. Glasberg and Brian C. J. Moore, "Derivation of auditory filter shapes from notched-noise data," *Hearing Research*, vol. 47, no. 1-2, pp. 103–138, August 1990.