Audio Engineering Society

Convention Paper

Presented at the 133rd Convention
2012 October 26–29 San Francisco, USA

DReaM: a novel system for joint source separation and multi-track coding

Sylvain Marchand[1], Roland Badeau[2], Cléo Baras[3], Laurent Daudet[4], Dominique Fourer[5], Laurent Girin[3], Stanislaw Gorlow[5], Antoine Liutkus[2], Jonathan Pinel[3], Gaël Richard[2], Nicolas Sturmel[3], Shuhua Zang[3]

[1] *Lab-STICC, CNRS, Univ. Western Brittany, Brest, France*

[2] *Institut Telecom, Telecom ParisTech, CNRS LTCI, Paris, France*

[3] *GIPSA-Lab, Grenoble-INP, Grenoble, France*

[4] *Institut Langevin, CNRS, ESPCI-ParisTech, Univ. Paris Diderot, Paris, France*

[5] *LaBRI, CNRS, Univ. Bordeaux 1, Talence, France*

Correspondence should be addressed to Sylvain Marchand (`Sylvain.Marchand@univ-brest.fr`)

ABSTRACT

Active listening consists in interacting with the music playing, has numerous applications from pedagogy to gaming, and involves advanced remixing processes such as generalized karaoke or respatialization. To get this new freedom, one might use the individual tracks that compose the mix. While multi-track formats loose backward compatibility with popular stereo formats and increase the file size, classic source separation from the stereo mix is not of sufficient quality. We propose a coder/decoder scheme for informed source separation. The coder determines the information necessary to recover the tracks and embeds it inaudibly in the mix, which is stereo and has a size comparable to the original. The decoder enhances the source separation with this information, enabling active listening.

1. INTRODUCTION

Active listening of music is both an artistic and technological topic of growing interest, that consists in giving to the music consumer 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 advanced remixing processes such as general-

ized karaoke (muting any musical element, not only the lead vocal track), adding effects on selected instruments, respatialization and upmixing. The applications are numerous, from learning/teaching of music to gaming, through new creative processes (disc jockeys, live performers, etc.).

To get this new freedom, a simple solution would be to give access to the individual tracks that compose the mix [1], 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 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 [2, 3].

In the DReaM project, we propose a system designed to perform source separation and accurately recover the separated tracks from the stereo mix. The system consists of a coder and a decoder.

The coder is used at the mixing stage, where the separated tracks are known. It determines the information necessary to recover the tracks from the mix and embeds it in the mix. In the case of PCM, this information is inaudibly hidden in the mix by a watermarking technique [4]. In the case of compressed audio formats, it can be embedded in a dedicated data channel or directly in the audio bitstream. With a legacy system, the coded stereo mix can be played and sounds just like the original, although some information is now included in it. Apart from this backward compatibility with legacy systems, an important point is the fact that the file size stays comparable to the one of the original mix, since the additional information sent to the decoder is rather negligible.

This decoder performs source separation of the mix with parameters given by the additional information. This Informed Source Separation (ISS) approach [5] permits to produce good separated tracks, thus enabling active listening applications.

The paper is organized as follows. Section 2 presents the DReaM project: its fundamentals and target applications. Section 3 introduces the mixing models we are considering, Section 4 describes the separation/unmixing methods we have developed so far in the project, and Section 5 illustrates the working prototypes available for demonstration purposes. Finally, Section 6 draws some conclusions and opens new perspectives.

2. THE DREAM PROJECT

DReaM[1] 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 with industrial finality, funded by the French National Research Agency (ANR). The project members are academics (LaBRI – University of Bordeaux, 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 [6]. The Lab-STICC – University of Brest will join the consortium, as the new affiliation of the first author and coordinator of the project. The Grenoble Innovation Alpes (GRAVIT) structure leads the technology transfer aspects of the project.

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 (generally 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(n un)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 separated. However, this multi-track approach is hardly feasible with a classic (stereophonic) audio CD.

Nowadays, the public 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*[2]. But

[1] see URL: `http://dream.labri.fr`
[2] see URL: `http://www.rockband.com`

in this case, to get the interaction the user has to buy the 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.

2.1. Project Goals and Objectives

Generally speaking, the project aims at solving an 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 observed mix.

On the one hand coding the solution (e.g., the individual tracks and the way how 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 (explained below). Indeed, the mixture signals should be realistic music pieces, ideally of professional quality, and the separation should be processed in real-time with reasonable computation costs, so that real-time sound manipulation and remixing can follow. 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 we propose 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 [7]. 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 Fig. 1), the extra information is estimated from the original source signals before the mixing process and is inaudibly embedded into the final mix. At the decoder (see Fig. 2), this information is extracted from the mix and used to assist the separation process. The residuals can be coded as well, even if joint coding is more efficient (not on the figures for the sake of simplicity, see Section 4 instead).

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)

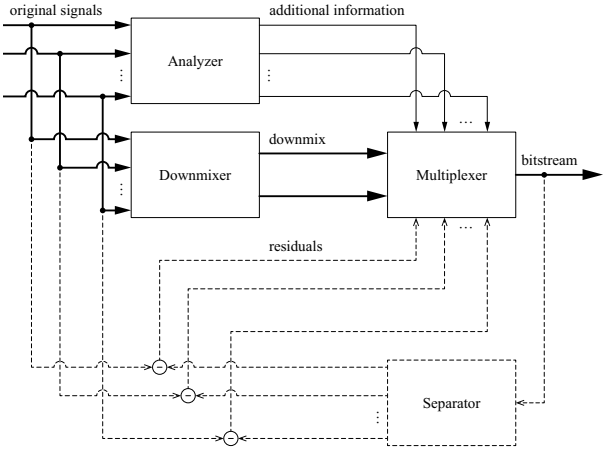

Fig. 1: General architecture of an ISS coder.

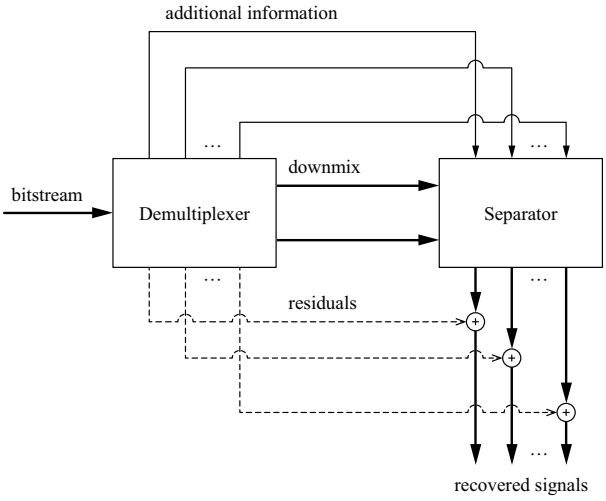

Fig. 2: General architecture of an ISS decoder.

2.1.1. From Active Audio CD...

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, see [4]) 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.

2.1.2. ...Towards Enhanced Compressed Mix

Now that the music is getting all digital, the consumer gets access to audio files instead of physical media. Although the previous strategy also applies to the (PCM) files extracted from the audio CD, most audio files are distributed in lossy compressed formats (e.g. ACC, MP3, or OGG).

We are currently working on the extension of the proposed techniques to compressed mixes, based on encouraging preliminary results [8]. The extra information can then either be included in some ancillary data, or be embedded (almost) inaudibly in the audio bitstream itself. The latter option is much more complicated, since lossy but perceptually lossless coding aims at removing inaudible information. Both coders (perceptual and informed) have then to be merged, to maintain a certain information trade-off.

2.2. Applications

Active listening [9] consists in 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 application of individual audio effects (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 [10], 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

Games, or "serious" games, can be very useful for music learning/teaching applications. 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 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).

2.3. ISS vs. SAOC

The DReaM project turns out to be close in spirit to the Remix system of Faller et al. [11]. We are also conscious that leaving artistic applications on uncompressed signals to more commercial applications on compressed formats now places the DReaM project next to MPEG Spatial Audio Object Coding (SAOC) [12], derived from the Spatial Audio Coding (SAC) approach of MPEG Surround (MPS) [13] and pioneering works on parametric multi-channel joint audio coding [14].

In MPS [13], perceptually relevant spatialization parameters such as interchannel loudness differences (ILD), interchannel time differences (ITD), and interchannel cross-correlations (ICC) are extracted from the multi-channel signal at the encoder. These parameters are transmitted to the decoder in addition to a mono/stereo downmix of the multi-channel signal. At the decoder, those parameters are used to respatialize the multi-channel audio scene from the downmix signal.

This approach has been extended later in SAOC [12] from the audio channels of the spatial image (acoustic scene) to audio objects (sound sources), opening new perspectives for active listening of music.

However, it must be noted that in contrast to SAC/SAOC, the goal of the ISS methods we propose (see Section 4 below) is from the beginning to completely separate the source signals and not only to resynthesize/respatialize the audio scene.

In particular, the spatialization parameters in SAC/SAOC are used to "redistribute" the content of spectral subbands of the downmix signal across the different output channels, but they cannot separate the contribution of two different sources that are present within the same subband (hence the sources are "respatialized together" and not clearly separated; e.g. see [14]). In contrast, the separation of two overlapping sources is precisely one of the original goals of our ISS methods. Note that some aspects of SAOC, notably the Enhanced SAOC option [15], tend to fill this gap by encoding additional information that achieves a (much) better separation of the audio objects. But this is done through separately encoding the residuals, which may be shown to be sub-optimal in terms of bitrate [7, 16], compared to a joint coding.

Finally, the connections between SAOC and DReaM might be stated this way: SAOC started from multi-channel coding and met source separation (using coding), whereas DReaM started from source separation and met coding.

3. THE MIXING MODELS

We present here the underlying model of all the methods we will consider in Section 4, as well as some generalizations.

We assume that the audio objects signals (or sources) are defined as M regularly sampled times series s_m of same length N. An audio object is thus understood in the following as a mono signal. Furthermore, we suppose that a mixing process produces a K-channel mixture $\{y_k\}_{k=1,\cdots,K}$ from the audio objects.

3.1. Linear Instantaneous Model

We first consider linear and time-invariant mixing systems. Formally, we suppose that each audio object s_m is mixed into each channel k through the use of some mixing coefficient a_{km}, thus:

$$y_k(t) = \sum_{m=1}^{M} y_{km}(t) \tag{1}$$

where

$$y_{km} = a_{km} \cdot s_m, \tag{2}$$

$\{y_{km}\}_{k=1,\cdots,K}$ being the (multi-channel) spatial image of the (mono) audio object s_m. In the stereo case where $K = 2$, we call this mono-to-stereo mixing.

We suppose that the mixing filters are all constant over time, thus leading to a time-invariant mixing system. We say that the mixing is linear instantaneous.

3.2. Convolutive Case

If the mixing coefficients a_{km} are replaced by filters, and the product in Eq. (2) is replaced by the convolution, we say that the mixing is convolutive. We can easily handle this case (see [17]) with the Short-Time Fourier Transform (STFT) representation if the length of the mixing filters is sufficiently short compared to the window length of the STFT, as:

$$Y_{km}(t,\omega) \approx A_{km}(\omega) S_m(t,\omega) \tag{3}$$

where $A_{km}(\omega)$ is understood as the frequency response of filter a_{km} at frequency ω. When the mixing process is linear instantaneous and time invariant, A_{km} is constant and the $K \times M$ matrix A is called the mixing matrix. When it is convolutive, this mixing matrix $A(\omega)$ is a function of ω. The mixing model can hence be written in the STFT representation as:

$$Y(t,\omega) \approx A(\omega) S(t,\omega) \tag{4}$$

where $Y = [Y_1, \cdots, Y_K]^\top$ and $S = [S_1, \cdots, S_M]^\top$ are column vectors respectively gathering all mixtures and sources at the time-frequency (TF) point (t,ω).

3.3. Non-linear Case

Of course, in real musical productions, non-linear effects such as dynamics compression are present in the mixing process. We have shown in [1] that it is possible to revert to the previous – linear – case by "moving" all the effects before the sum operation of the mixing model. The problem with this approach is that it might lead to "altered" sound objects – i.e. "contaminated" by the effects – and thus harder to use for some active listening scenarios without noticeable artifacts. Another approach is to invert the effects in order to revert to the linear case. This is clearly out of the scope of this paper, where we rather focus on the inversion of the sum operation of the mixing model, in order to estimate the original sources. However, the methods presented in the next section have proved to be quite resistant to non-linearities of the mixing process.

3.4. Image-Based Model

In real-world conditions, the mixing process may be much harder to model [1]. Take for instance the stereo sub-mix of a multi-channel captured drum set, or the stereo MS recording of a grand piano. Then the solution is to not consider audio objects anymore but rather directly their spatial images. Source separation consists then in inverting the sum of Eq. (1), to recover the M separate images $\{y_{km}\}_{k=1,\cdots,K}$ from the mixture $\{y_k\}_{k=1,\cdots,K}$. Each image has then the exact number of channels as the mix ($K = 2$ for a stereo mix). Such model will be referred to as stereo-to-stereo mixing. In such case, audio objects are not separated, but the modification of the separated images can still allow a substantial amount of active listening scenarios, including remixing and generalized karaoke. Respatialization, however, can be more difficult.

4. INFORMED SEPARATION METHODS

The objective of informed source separation is hence to compute some additional information that allows to recover estimates of the sources given the mixture $\{y_k\}_{k=1,\cdots,K}$. Depending on the method, these sources can be either the audio objects s_m or their spatial images $\{y_{km}\}_{k=1,\cdots,K}$ ($K = 2$ for stereo).

For the computation of the additional information, we assume that s_m and A are all available at the coder stage. Of course, the main challenge is to develop techniques that produce good estimates with an additional information significantly smaller than the one needed to directly transmit s_m.

Over the past years, we already proposed several informed source separation methods. 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 paper. The detailed descriptions of the methods can rather be found in [18], [19], [20], and [21], while their comparison is done in [17].

4.1. Time-Frequency Decomposition

All the methods we propose are based on some time-frequency (TF) decomposition, either the MDCT or the STFT, the former providing critical sampling and the latter being preferred for the mixing model (see Section 3) and for filtering thanks to the convolution theorem.

Then, for each TF point, we determine the contribution of each source using several approaches and some additional information.

4.2. Additional Information

In the following, we assume that the encoder is provided with the knowledge of the mixing matrix A. However, this matrix may be estimated as demonstrated in [19]. This information may be used either directly or by deriving the spatial distribution of the sources. Then, our different methods have specific requirements in terms of additional information.

4.2.1. Source Indices

The first information we used was the indices of the two most prominent sources, that is the two sources with the highest energy at the considered TF point. As explained below, this information can be used to solve the interference of the sources at this point. This information can efficiently be coded with $\lceil \log(M(M-1)/2) \rceil$ bits per TF point.

4.2.2. Source Energies

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. This information can efficiently be coded using for example the Equivalent Rectangular Bandwidth (ERB) and decibel (dB) scales, closer to the perception, together with entropy coding [20],

or alternatively with Non-negative Tensor Factorization (NTF) techniques, as demonstrated in [19, 16].

4.3. Several Approaches

The majority of our ISS methods aims at extracting the contribution of each source from each TF point of the mix, at least in terms of magnitude, and of phase too for most of the methods.

Our first method performs a local inversion [18] 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. As discussed below, this method gives the best results with the Signal-to-Distortion Ratio (SDR) objective measure of BSS-Eval [22]. But the problem is that the remaining $M-2$ sources exhibit a spectral hole (no estimated signal), which is perceived as quite annoying in subjective listening tests [20]. Also, this method requires the mixing matrix A to be of rank M.

Our second method performs Minimum Mean-Square Error (MMSE) filtering [19] 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 NTF or image compression techniques. Although this method produces results with a lower SDR, the perceived quality is higher, which matters to the listener. 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 M.

Our third method performs linearly constrained spatial filtering [20] 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). In the stereo case ($K=2$), 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 best estimates than the early MMSE approach, at least with the PEMO-Q [23] measure, closer to the perception.

Our fourth method performs iterative phase reconstruction and is called IRISS (Iterative Reconstruction for Informed Source Separation) [21]. 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 Eq. (3) is satisfied at each iteration very much like the Multiple Input Spectrogram Inversion (MISI) method [24]. 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 is specifically designed for mono mixtures ($K=1$), where it gives the best results, and does not yet benefit from the case $K>1$.

The main remaining issue with the aforementioned methods is that their performance is bounded. Other methods recently proposed [7, 16] are based on source coding principles in the posterior distribution of the sources given the mixtures and should permit to reach arbitrary quality provided that the bitrate of the additional information is sufficient.

4.4. Performances

The quality performance of the system now reaches the needs of many real-life applications (e.g. industrial prototypes, see Section 5 below) with ongoing technology transfers and patents. The comparison of the current implementation of our four methods can be found in [17], for the linear instantaneous and convolutive cases (see Section 3), using either the objective SDR criterion of BSSEval [22] or the PEMO-Q measure [23], closer to perception. It turns out that the first method (local inversion) exhibits the best SDR (objective) results, while the third method (constrained spatial filtering) exhibits

the best PEMO-Q (more subjective) scores; this was also verified in a formal listening test [20]. It is important to note that the complexity of these methods is low, enabling active listening in real time. Moreover, as shown in [17], the typical bitrates for the additional information are approximately 5-10kbps per mixture and audio object, which is quite reasonable.

5. PROTOTYPES

Multiple versions of the DReaM system allow applications to uncompressed (PCM) and compressed (AAC/MP3/OGG) mixdown with mono-to-mono, mono-to-stereo, and stereo-to-stereo mixtures including artistic effects on the stereo mix [1].

5.1. DReaM-RetroSpat

We have presented in [25] a real-time system for musical interaction from stereo files, fully backward-compatible with standard audio CDs (see Fig. 3). This system manages the mono-to-stereo case and consists of a source separator based on the first DReaM method of Section 4 (local inversion) and a spatializer, RetroSpat [26], based on a simplified model of the Head-Related Transfer Functions (HRTF), generalized to any multi-loudspeaker configuration using a transaural technique for the best pair of loudspeakers for each sound source. Although this quite simple technique does not compete with the 3D accuracy of Ambisonics or holophony (Wave Field Synthesis – WFS), it is very flexible (no specific loudspeaker configuration) and suitable for a large audience (no hot-spot effect) with sufficient quality.

The resulting software system is able to separate 5-source stereo mixtures (read from audio CDs or 16-bit PCM files) in real time and it enables the user to remix the piece of music during playback with basic functions such as volume and spatialization control. The system has been demonstrated in several countries with excellent feedback from the users/listeners, with a clear potential in terms of musical creativity, pedagogy, and entertainment.

5.2. DReaM-AudioActivity

The DReaM-AudioActivity prototype (see Fig. 4) targets consumer/prosumer applications of the ISS technologies issued of DReaM. The software is written in such as way that each separation method can be included as a separate `C++` subclass, but at the time of writing of this article, only the MMSE filter

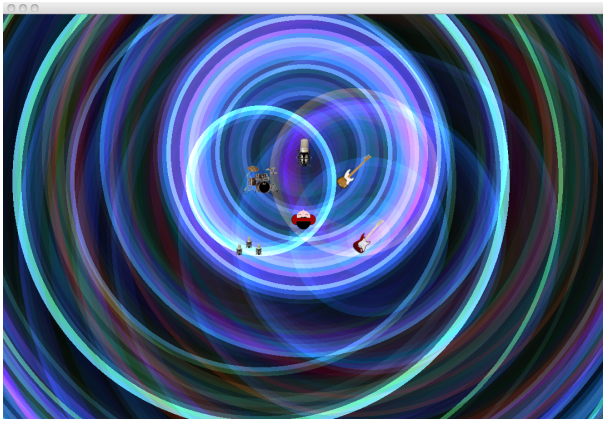

Fig. 3: From the stereo mix, the DReaM-RetroSpat player permits the listener (center) to manipulate 5 sources in the acoustic space (and to visualize the sound propagation).

method was implemented. This work is supported by GRAVIT in collaboration with the DReaM team.

This prototype addresses the issue of studio music production, that is the stereo-to-stereo case. In some cases, the mix may not even be the exact sum of the stereo sources: dynamics processing can be applied and estimated *a posteriori* [1]. The coder performs, in almost real time, high-capacity watermarking of the separation information from the separated stereo tracks into the artistic mix coded in 16-bit PCM. The decoder performs offline reading of this watermark and performs the separation and re-mixing in real time. The number of tracks than can be included in the mix is only limited by the capacity of the watermark. Vector optimization of the audio processing core gives very low CPU usage during live separation and remixing. The end-user can then modify the volume and stereo panning of each source in real time during playback. Automation of global and per track volume and panoramic is also possible. As always, the coded mix is backward compatible with standard 16-bit PCM playback software programs with little to no audio quality impact.

6. CONCLUSION

In this paper, we have presented the DReaM project. Originally thought as a way to interact with the music signal through its real-time decomposition/manipulation/recomposition, the emphasis has

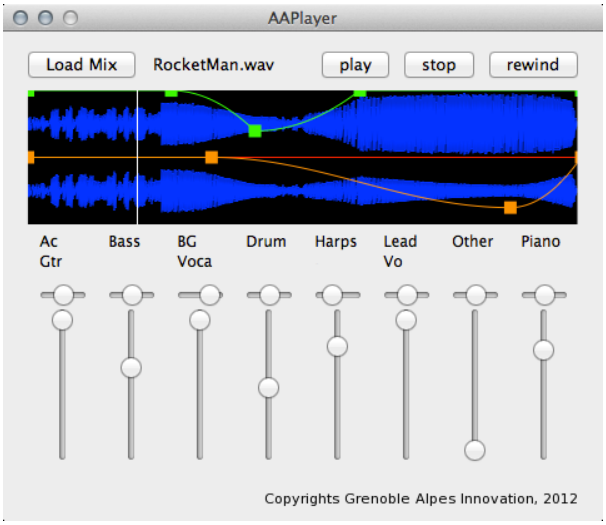

Fig. 4: Manipulation of a 8-source mix by the DReaM-AudioActivity player.

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. Some of our techniques have been implemented in software prototypes, for demonstration purposes. These prototypes enable the user to perform, for instance, generalized karaoke and respatialization. We are currently extending our methods to compressed audio formats. We propose to compare our approach to e.g. MPEG SAOC in the near future, and envisage generalizing this informed approach to other problems than source separation, e.g. to the inversion of audio effects.

7. ACKNOWLEDGMENTS

This research was partly supported by the French ANR *(Agence Nationale de la Recherche)*, within the scope of the DReaM project (ANR-09-CORD-006).

8. REFERENCES

[1] N. Sturmel, A. Liutkus, J. Pinel, L. Girin, S. Marchand, G. Richard, R. Badeau, and L. Daudet, "Linear mixing models for active listening of music productions in realistic studio conditions," in *Proceedings of the 132nd AES Convention*, Budapest, Hungary, April 2012.

[2] P. Comon and C. Jutten, Eds., *Handbook of blind source separation – Independent component analysis and applications*, Academic Press, 2010.

[3] A. Ozerov and C. Févotte, "Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 3, pp. 550–563, 2010.

[4] J. Pinel, L. Girin, C. Baras, and M. Parvaix, "A high-capacity watermarking technique for audio signals based on MDCT-domain quantization," in *Proceedings of the International Congress on Acoustics (ICA)*, Sydney, Australia, August 2010.

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

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

[7] A. Ozerov, A. Liutkus, R. Badeau, and G. Richard, "Informed source separation: source coding meets source separation," in *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, New York, USA, October 2011, pp. 257–260.

[8] L. Girin and J. Pinel, "Informed audio source separation from compressed linear stereo mixtures," in *Proceedings of the 42nd AES Conference*, Ilmenau, Germany, July 2011.

[9] P. Lepain, *Recherche et applications en informatique musicale*, chapter Écoute interactive des documents musicaux numériques, pp. 209–226, Hermes, Paris, France, 1998, In French.

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

[11] C. Faller, A. Favrot, J.-W. Jung, and H.-O. Oh, "Enhancing stereo audio with remix capability," in *Proceedings of the 129th AES Convention*, San Francisco, California, USA, November 2010.

[12] J. Engdegård, C. Falch, O. Hellmuth, J. Herre, J. Hilpert, A. Hölzer, J. Koppens, H. Mundt, H. Oh, H. Purnhagen, B. Resch, L. Terentiev, M. Valero, and L. Villemoes, "MPEG spatial audio object coding – the ISO/MPEG standard for efficient coding of interactive audio scenes," in *Proceedings of the 129th AES Convention*, San Francisco, California, USA, November 2010.

[13] J. Herre, K. Kjörling, J. Breebaart, C. Faller, S. Disch, H. Purnhagen, J. Koppens, J. Hilpert, J. Rödén, W. Oomen, K. Linzmeier, and K. Chong, "MPEG surround – the ISO/MPEG standard for efficient and compatible multichannel audio coding," *Journal of the AES*, vol. 56, no. 11, pp. 932–955, November 2008.

[14] C. Faller, "Parametric joint-coding of audio sources," in *Proceedings of the 120th AES Convention*, Paris, France, May 2006.

[15] C. Falch, L. Terentiev, and J. Herre, "Spatial audio object coding with enhanced audio object separation," in *Proceedings of the International Conference on Digital Audio Effects (DAFx)*, Graz, Austria, September 2010, pp. 1–7.

[16] A. Liutkus, A. Ozerov, R. Badeau, and G. Richard, "Spatial coding-based informed source separation," in *Proceedings of the European Signal Processing Conference (EUSIPCO)*, Bucharest, Romania, August 2012.

[17] A. Liutkus, S. Gorlow, N. Sturmel, S. Zhang, L. Girin, R. Badeau, L. Daudet, S. Marchand, and G. Richard, "Informed audio source separation: a comparative study," in *Proceedings of the European Signal Processing Conference (EUSIPCO)*, Bucharest, Romania, August 2012.

[18] M. Parvaix and L. Girin, "Informed source separation of linear instantaneous underdetermined audio mixtures by source index embedding," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 6, pp. 1721–1733, 2011.

[19] A. Liutkus, J. Pinel, R. Badeau, L. Girin, and G. Richard, "Informed source separation through spectrogram coding and data embedding," *Signal Processing*, vol. 92, no. 8, pp. 1937–1949, 2012.

[20] S. Gorlow and S. Marchand, "Informed audio source separation using linearly constrained spatial filters," *IEEE Transactions on Audio, Speech, and Language Processing*, 2012, In Press.

[21] N. Sturmel and L. Daudet, "Informed source separation using iterative reconstruction," *IEEE Transactions on Audio, Speech, and Language Processing*, 2012, In Press.

[22] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1462–1469, 2006.

[23] R. Huber and B. 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, 2006.

[24] D. Gunawan and D. 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.

[25] S. Marchand, B. Mansencal, and L. Girin, "Interactive music with active audio CDs," *Lecture Notes in Computer Science – Exploring Music Contents*, vol. 6684, pp. 31–50, August 2011.

[26] J. Mouba, S. Marchand, B. Mansencal, and J.-M. Rivet, "RetroSpat: a perception-based system for semi-automatic diffusion of acousmatic music," in *Proceedings of the Sound and Music Computing (SMC) Conference*, Berlin, Germany, July/August 2008, pp. 33–40.