Linear Mixing Models for Active Listening of Music Productions in Realistic Studio Conditions

Nicolas Sturmel, Antoine Liutkus, Jonathan Pinel, Laurent Girin, Sylvain Marchand, Gael Richard, Roland Badeau, Laurent Daudet

Audio Engineering Society

Convention Paper

Presented at the 132nd Convention
2012 April 26–29 Budapest, Hungary

Linear mixing models for active listening of music productions in realistic studio conditions

Nicolas Sturmel[1], Antoine Liutkus[3], Jonathan Pinel[2], Laurent Girin[2], Sylvain Marchand[4], Gaël Richard[3], Roland Badeau[3], Laurent Daudet[1]

[1] *Institut Langevin, CNRS, ESPCI-ParisTech, Université Paris Diderot, 75005 PARIS, FRANCE*

[2] *GIPSA-Lab, Grenoble-INP, GRENOBLE, FRANCE*

[3] *Institut Telecom, Telecom ParisTech, CNRS LTCI, PARIS, FRANCE*

[4] *Université de Bretagne Occidentale, BREST, FRANCE*

Correspondence should be addressed to Nicolas Sturmel (`nicolas.sturmel@espci.fr`)

ABSTRACT

The mixing/demixing of audio signals as addressed in the signal processing literature (the "source separation" problem) and the music production in studio remain quite separated worlds. Scientific audio scene analysis rather focuses on "natural" mixtures and most often uses linear (convolutive) models of point sources placed in the same acoustic space. In contrast, the sound engineer can mix musical signals of very different nature and belonging to different acoustic spaces, and exploits many audio effects including non-linear processes. In the present paper we discuss these differences within the strongly emerging framework of active music listening, which is precisely at the crossroads of these two worlds: it consists in giving to the listener the ability to manipulate the different musical sources while listening to a musical piece. We propose a model that allows the description of a general studio mixing process as a linear stationary process of "generalized source image signals" considered as individual tracks. Such a model can be used to allow the recovery of the isolated tracks while preserving the professional sound quality of the mixture. A simple addition of these recovered tracks enables the end-user to recover the full-quality stereo mix, while these tracks can also be used for, e.g., basic remix / karaoke / soloing and re-orchestration applications.

1. INTRODUCTION

Active listening consists in performing various operations that modify the elements and structure of the music signal during the listening of a music piece. This process, often simplistically called remixing, includes generalized karaoke (music minus one: abil-

ity to suppress an instrument), re-spatialization, 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 personalization of the musical piece through various re-orchestration techniques. Alternately, active listening solutions intrinsically provide simple frameworks to the artists to produce different artistic versions of a given piece of music. Moreover, it is an amazing framework for music learning/teaching applications.

Active listening applications have received a growing attention in the past years, as illustrated by multi-track formats such as iKlax [5] or MXP4[1], musical games such as Harmonix Rock Band[2], and objects-oriented audio standards such as MPEG-SAOC [3]. Those technologies all benefit from the prior recording and processing of the separate elements. Indeed, in order to achieve active listening, one has to control the so-called "stems" within the mixture. A stem is a signal that represents a track, an instrument or a group of instruments that have to be processed together according to some (arbitrary) artistic criterion. For example, the drums, which are a combination of several percussive instruments, can be considered as a single stem if the complete drums set is to be controlled globally, whereas it can be decomposed into several stems, e.g., for pedagogical applications.

In active listening, a stem plays the role of what is referred to as "source signal" in the signal processing literature. Because the stems have to be considered at both the music production level (the recording and mixing studios) and at the user level (personal music player), an active listening system has the form of a coder/decoder system, as illustrated on Figure 1. The coding stage allows direct or indirect transmission of the source signals and the decoding stage allows recovery, individual manipulation, and remixing of these source signals. The simplest case is the multi-track format (Figure 1a): in this case, the full original source signals are perfectly known at the decoder. The problem here is that a very limited number of commercial songs are distributed in this format. The size of the multi-track files and the reluctance of the music industry to give unlimited access to the separated stems are probably the most important limitations of such distribution formats.

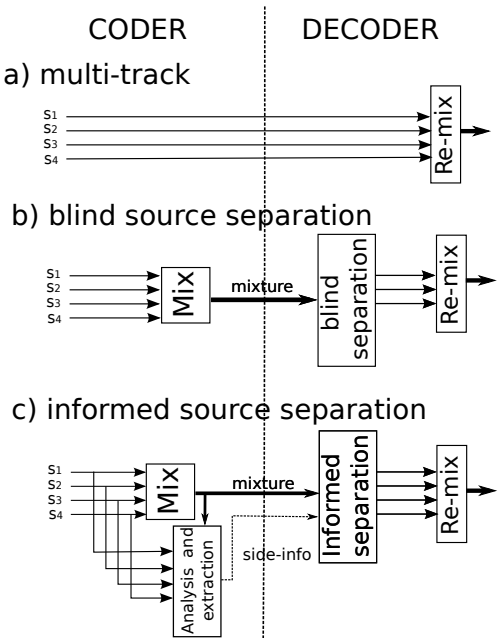

Fig. 1: Coder/Decoder schemes for active listening.

Most often, only the mix signal is available at the decoder. Source separation may then be used to recover the source signals (Figure 1b). Here, the term source separation refers to the process of recovering the source signals from the mix signal only. This includes different approaches [8]. However, despite of the intensive efforts of the research community in this topic in the last decades, these "blind" source separation approaches still do not accurately recover the original source signals for real-world complex audio mixtures. The quality of separated source signals is thus generally not sufficient for active listening applications. In particular, it is not guaranteed to estimate the correct number of sources as shown on the figure.

Recent approaches try to draw a line between multi-track (i.e. source coding) and source separation, merging these two aspects in a hybrid approach: Informed Source Separation (ISS) [4, 14, 13, 9, 6, 12] and Audio Object Coding (AOC) [2, 3, 7] consist in extracting a prior knowledge from the signals at the coder stage to facilitate the separation at the decoder stage (Figure 1c). This knowledge is

[1] http://www.mxp4.com/
[2] http://www.harmonixmusic.com/

compressed and transmitted to the decoder as side-information, either in a separate channel, or embedded within the mix signal bitstream, or hidden within the mix signal samples by watermarking techniques. The major advantages of this approach is that the music signal is provided with a format that is totally compliant with usual music players (mostly PCM or compressed format), so that the default "passive" listening can be performed on any player. On top of that, the side-information is usually lower than the compressed versions of the separated signals that would be transmitted with the mix.

In all cases, the quality of mix and remix is of paramount importance in commercial music: mixing is not straightforward. In a typical studio setup, various non-linear and non-instantaneous effects are used at different stages of the production chain. This raises two issues for active listening applications:

1. If one can recover the separated signals, do they take into account full or part of these mixing effects? And thus, which part of the effects remains in charge of the remix?
2. In the case where source separation is used to provide the signals, how are these effects taken into account in the separation process?

So far, these two issues have been poorly addressed, if not avoided. At the end of the music production chain, mixing and remixing are often reduced in the audio processing literature to a simple Linear Instantaneous Stationary (LIS) process, which does not provide the full flexibility of studio effects. In other words, the LIS model does not apply in the case of artistic music (re)mixing. In the case of audio source separation, most of the literature addresses linear, instantaneous or convolutive, mixtures but non-linear mixture analysis remains marginal[3].

As will be presented later, the studio constraints are not appropriate for simple and efficient source separation methods based on this linearity assumption. The goal of this paper is precisely to clarify the links between studio mixing techniques and demixing/remixing models, as used in audio scene analysis and source separation techniques, within the active listening framework. In particular, an effort is done on the disambiguation of the terms source, track, stem and signal in relation to the problem. This paper also presents a generalized linear mixing model that conciliates the studio production constraints and the efficiency of some existing separation and remixing methods based on the LIS assumption. Note that, because ISS and AOC allow the access to the (different steps of) source/mix processing, they offer a privileged framework for the present study. Some considerations may thus be specifically applicable to ISS/AOC systems, but some others may concern the whole source separation framework.

This paper is organized as follows. In Section 2 we briefly present the fundamental models of audio source mixture and separation as generally considered in the literature. In Section 3 we present a typical studio mixing setup, as generally implemented on Digital Audio Workstations (DAW). In Section 4, we detail the differences between these two frameworks, and underline the difficulties, if not impossibilities, of directly applying usual mixture models to music produced in studio. In Section 5 we then extend the "studio process" to a distributed instantaneous form applicable to existing active listening systems in real conditions, using tools already available in professional music production. Section 6 concludes the paper and opens on future works.

2. A BRIEF REVIEW OF MIXTURE MODELS

As seen before, the most simple mixing model is the LIS process, which involves only one invariant mixing parameter per source per channel:

$$m_j(n) = \sum_i a_{i,j} s_i(n), \quad (1)$$

where m_j is the mixture signal on output channel j, s_i are the source signals and $a_{i,j}$ is the mixing coefficient of source i onto channel j. Such mixture is very simple but has poor physical reality in the case of sounds (a simple "pan-pot"). It is however often chosen because of its linearity and the small number of mixing coefficients.

More complex models involve the observation of an acoustical scene [15] where the sources are recorded

[3] An example of "post non-linear" configuration can be found in [16], but the mix process before the non-linear transform is limited to instantaneous and determined, a quite unusual configuration in studio mixing.

using multiple microphones. Often, the number of channels J is 2 as in the case of stereophonic sounds. This is directly linked to the fact that humans perceive sounds with two ears. This model leads to more complex mixtures such as the linear convolutive model. For each source and each microphone, an impulse response $r_{i,j}(n)$ that depends on their absolute position in space, can be computed so that the mixture can be modeled as:

$$m_j(n) = \sum_i \sum_{l=0}^{\infty} r_{i,j}(l) s_i(n-l). \quad (2)$$

The linear instantaneous model and, more importantly, the linear convolutive model are the basis for a large amount of work in source separation of "real-life" audio scene (see a review in, e.g., [10]). However, these models are very limiting in regards to the various possibilities of professional music mixing (and also demixing as long as active listening from the mix signal is involved) because they only consider linear processing of point sources all placed in the same acoustical space. However, these models have the advantage of being very simple and tightly linked to the way the human brain listens to music. These models also offer a privileged framework in the case of videoconference and robot audition because of the unique and well defined acoustic space of such applications.

3. A TYPICAL DAW MIXING SETUP

Let us consider a typical DAW mixing desk used for the production of professional-quality music from individually recorded tracks, with arbitrary audio effects. Note that the notion of source is irrelevant here: tracks are the elements that are processed during mixing. One can classify effects in three categories:

1. Linear instantaneous effects: gain and panning (different gains for different channels)
2. Linear convolutive effects: equalization, reverberation, delay...
3. Non-linear effects: distortion, chorus, dynamic processing and various complex signal processing such as denoising or non-linear analog modeling.

A typical DAW setup is presented on Figure 2 for a conventional stereo 2-channel mix. Previously recorded tracks are considered as the inputs of the system. Note that without loss of generality, auxiliary mixing busses (effects send, sub-mixes) are not presented on the figure: they are only specific cases of this general overview. The general process can be sequenced as follows. The listed effects are first applied on a per-track basis, with mono or stereo tracks, between step 1 (tracks, t_i) and step 2 (tracks with effects). The mono tracks are then panned between left and right channels with simple gains or more sophisticated effects to obtain spatial images. Stereo effects may be correlated from one channel to another (last stereo channel of Figure 2). At step 3, each track has been processed to its multi-channel version $t_{i,j}$. These multi channels versions are then summed to provide the so-called "master" (step 4). The master bus is then processed, with convolutive and non-linear effects. Those additional effects lead to the so called "artistic mix" or "commercial mix" (step 5), the final product experienced by the end-user. In summary, considering a per track mixing function $N_{i,j}[.]$ and a master processing function $O_j[.]$, the mixture m on channel j is given by:

$$m_j = O_j\left[\sum_i N_{i,j}(t_i(n))\right] = O_j\left[\sum_i t_{i,j}(n)\right]. \quad (3)$$

Note that between steps 4 and 5, only few effects are present. Generally only equalization, dynamic processing and sometimes reverberation are applied. Non-linear effects other than dynamic processing on the master track are rare, but this dynamic processing is generally of great importance. For example, it is used to modify the mixture so that it fits the distribution medium (e.g., "loud" version for radio broadcasting, see also the loudness war problem [17]).

4. LINK BETWEEN SIGNAL PROCESSING MODELS AND STUDIO REALITY

As one can see from the two preceding sections, the difference between classical mixture modeling and practical mixing in music production is significant. The present Section discusses the limitations of the existing models with regards to the music production practices. Different existing implementations will also be discussed.

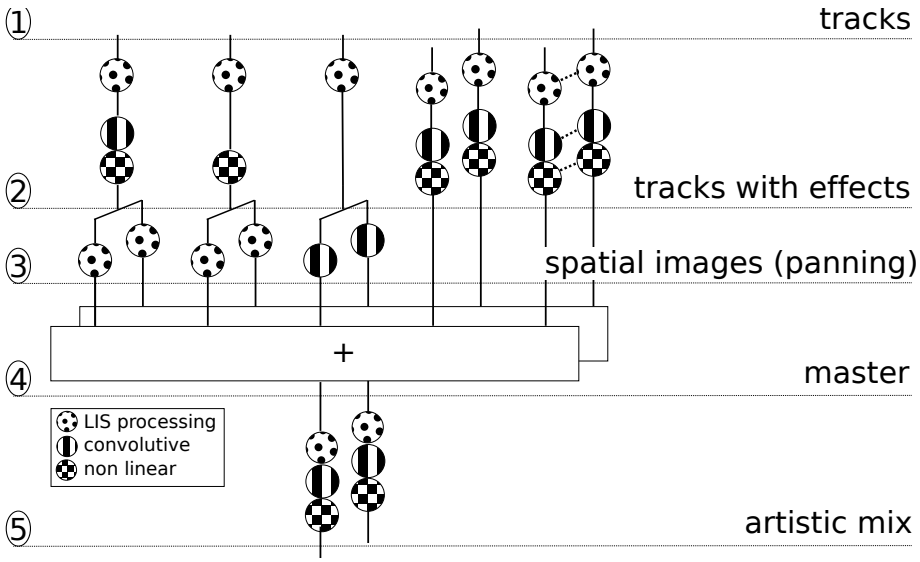

Fig. 2: A classical DAW 2-channel setup with mono and stereo sources. Circles indicate arbitrary effects processing.

4.1. Source images

Consider a set of "tracks" used for mixing. Thanks to studio practices (e.g. close miking, acoustic barriers, re-recording) separation between tracks is often excellent. The basic idea of active listening is then to capture the separate tracks t_i (stage 1 of Figure 2) and give the end user the ability to modify them via a mixing desk. However, some of these tracks capture (a part of) the same instrument (e.g. drums, piano) or the same group of instrument (e.g. : choir, brass section). The work of a mixing engineer often consists in assembling these tracks into consistent stereophonic (or multichannel) submixes. Take for instance a drums kit captured with 12 microphones, the corresponding tracks are assembled to a consistent stereophonic submix.

Actually, the mixing engineer tries to "build" an image of each instrument. When listening to the mix, the brain of the auditor then decomposes the mix into these images [1], separating the different so-called "source images" [11, 18]. Active listening systems must then take this constraint into account:

- Give access to the separated tracks but with a symbolic link between tracks related to the same musical image.
- Directly give access to the source images as composed by the engineer, rendering this symbolic link implicit.

In all cases, the end user gets to modify each (or a selected number of) source images composing the mix. Note that the term "source image" is ubiquitous as it may refer to an ensemble (e.g. choir), an instrument (e.g. piano, drums) or a specific acoustically separable part of an instrument (e.g. snare drum). Each source image is arbitrarily defined according to its potential use at the active listening stage. Note that the separation quality may be impacted by the acoustical separation of the recordings.

We then define the kth source image $s_{k,j}$ on channel j contained in the mixture m_j. Source images are obtained at level 3 by assembling the processed tracks $t_{i,j}$ in different sets. Let us designate one set by $\mathcal{I}_k$, then each track i is contained in one and only one set $\mathcal{I}_k$, and we have:

$$s_{k,j}(n) = \sum_{i \in \mathcal{I}_k} t_{i,j}(n). \tag{4}$$

Note that, as expected, source images are multichannel versions of the sources s_i, but the former are practical representations whereas the latter are ideal representations. We define the mix as a sum of source images $s_{k,j}$ captured as a set of multichannel

tracks from the level 3 of a DAW mixing desk :

$$\tilde{m}_j(n) = \sum_k s_{k,j}(n) = \sum_k \left[\sum_{i \in \mathcal{I}_k} t_{i,j}(n) \right]. \qquad (5)$$

If there exists a (physical) link between the channels at the signal production level, or at the mixing level, then there may be an identifiable relation within the source images, i.e. between $s_{k,1}$ and $s_{k,2}$ for a 2-channel mix. This relation may be exploited in the demix/remix application [4].

4.2. Inverting the mixing effects

Simple mixing models, as presented in Equation (2), only consider the (idealized) source s_i and not its (practical) source image artistically constructed by the sound engineer. In order to take into account the real mixing condition, one could define a per source mixing function $\beta_{i,j}$ that changes each ideal source s_i into its image $s_{i,j}$ on every channel (level 3 of Figure 2) so that the raw mix $\tilde{m}_j$ is given by:

$$\tilde{m}_j(n) = \sum_i \beta_{i,j}\left(s_i(n)\right). \qquad (6)$$

Active listening is then done by inverting or modifying $\beta_{i,j}$, but this raises various issues:

1. Effects used during mixing are often complex and even non linear. They are therefore difficult to invert.
2. During the mix, some processing is done to enhance the coherence between tracks that will build a common source image. Inverting such processing would break this coherence.
3. If the instrument is large (e.g. piano, choir, or drums) it might be intrinsically defined as a source image (e.g. using stereo capture).

Note that the difference between Equations (6) and (5) is based on the inversion of the mixing process. Therefore, the channel-based approach of Equation (5) is more general in the case of artistic mixes. The main drawback of the channel-based approach is that using signals that already carry their convolutive term and panning effects may notably limit the possibility of re-spatialization. But even so, it can be reasonably argued that inversion of spatialization is expected to be much easier on a single well separated source signal than on a complete mix signal. In the case of ISS, a representation of this spatialization function could very well be embedded within the mix to facilitate its inversion.

4.3. Master effects

As presented before, the use of source image may be the simplest choice for active listening. Practically speaking, the engineer has only to "solo" the tracks corresponding to a selected source image set $\mathcal{I}_k$ in order to record it separately. However, the presence of effects on the master may be problematic, especially if they are non linear. Such effects are modeled by the term O of Equation (3). Take for instance the scheme of Figure 2: the rough mix is often dynamically processed to make it "louder" (additional reverberation and equalization can also be applied). Extreme dynamic processing (also known as brick-wall limiting) is also commonly used to cut the signal above a certain threshold. Such highly non-linear dynamic processing can produce additional spectral content on the mix signal and can even change the spatial perception of the sound. But these modifications are not present on the source images as captured at level 3 of Figure 2, since they are captured before the summing stage. Therefore, at the decoder of an active listening system, the summation of individual/separated source image signals, as they appear before the master processing, cannot give back the full artistic properties of the musical piece.

4.4. Limitations of the existing techniques

The use of multi-track format (Figure 1a) taken at stage 3 of Figure 2 is prejudicial to the global artistical quality of the reconstructed mix. Because the end-user has not access to the processing done on the master, some of the artistic quality of the mixing is lost. Moreover, trying to subtract a source image from the artistic mix might not allow full quality "music minus one" applications because of these added master effects.

Since source separation (Figure 1b) relies on knowledge of the final mixture (where the master effects are present), reconstructed source images may contain part of these effects: the main idea behind source separation is that the error between the sum of the estimated source images and the original mix

is zero. Then, the spectral content added by additional processing would anyway be distributed onto the reconstructed source images regardless of their capture point on Figure 2. However this distribution is not well controlled. This has been observed in SAOC [3], ISS [9] and blind separation [11].

In contrast, the use of an informed approach (Figure 1c) can allow a better control of this problem. We focus on this point on the next section.

5. GENERAL SEPARATED MIXING MODEL

After the discussion in the previous section, it appears that the remaining important question is how to allow the processing effects on the master between steps 4 and 5 to be distributed on each source image. This section presents a new model that offers versatile possibilities for the implementation of source separation methods. In particular we propose a targeted linearization of the dynamic processing (including all kinds of compression and limiting) so that we can reduce the artistic mix to a sum of what will be presented as "generalized source images".

5.1. Back to linear: Distributing the dynamic processing effects

Let us remind that the processed track signal i at level 3 of Figure 2 is given by $t_{i,j}(n)$. As mentioned before, two kinds of effects can be applied to the master:

- $c_j(n)$ represents a convolution process that encompasses all linear time-invariant processing (equalization, reverberation) on the master.
- $n_j(.)$ is a non-linear function at the end of the processing chain (mainly modeling dynamic processing, see below).

The master signal on channel j is thus given by:

$$m_j(n) = O_j\left(\sum_i t_{i,j}(n)\right) = n_j\left(c_j(n) * \sum_i t_{i,j}(n)\right). \quad (7)$$

This model represents then the complete mixing process. The objective is here to transform this process into an equivalent linear process. For this aim, the convolutive process $c_j(.)$ can be first easily distributed to each pre-master track to provide a new convolved track $c_j(n) * t_{i,j}(n)$. The non-linear term $n_j(.)$ is more problematic at first sight. However, although non-linear effects are various in studio, only a few of them are actually used on busses of the mixing desk. Most of the non-linear effects are dynamic processors such as compressors or limiters. This is especially true for the master bus: as mentioned before, in most conventional mixing, $n_j(.)$ represents the dynamic processing only, and we focus on this effect in the following.

Dynamic processing is composed of two chained components, as represented on the top of Figure 3: the dynamic detection and the gain (reduction). Dynamic detection consists in estimating the instantaneous gain $g_j(n)$ from the input mix $\hat{m}_j(n) = \sum_i c_j(n) * t_{i,j}(n)$. The gain chain consists in applying this gain to the input mix signal as a simple time-varying envelope to obtain the final mix signal $m_j(n) = g_j(n)\hat{m}_j(n)$. At this point, it is of primary importance to note that dynamic processing is a non-linear process from the "control" signal point of view, but it is a linear (non time-invariant) process from the "target" signal point of view, i.e. the signal on which the dynamic compression is applied. In other words, the gain $g_j(n)$ can be distributed on each convolved track signal, so that:

$$m_j(n) = \sum_i g_j(n)(c_j(n) * t_{i,j}(n)).$$

As opposed to other non-linear effects, dynamic processing with a side chaining input can be processed as if it were linear. This way, we are able to compute the spectral modification induced by the dynamic processing on each track. We can thus redefine the track signals at the final master level as:

$$\tilde{t}_{i,j}(n) = g_j(n)(c_j(n) * t_{i,j}(n)),$$

and thus we have

$$m_j(n) = \sum_i \tilde{t}_{i,j}(n).$$

Thanks to the linearity of Equation 4, all the previous considerations can also be applied on the source images. Therefore we can introduce the "generalized" source image $\tilde{s}_{k,j}$ given by:

$$\tilde{s}_{k,j}(n) = g_j(n)(c_j(n) * s_{k,j}(n)) = \sum_{i \in \mathcal{I}_k} \tilde{t}_{i,j}(n),$$

and the final master can be redefined as a linear mixture of generalized source images:

$$m_j(n) = \sum_k \tilde{s}_{k,j}(n).$$

Of course, in such a mixture, the relation between the images of the same source signal within the different channels may not be characterized/identified easily, depending on the nature of the processes at the pre-mix and post-mix levels. Therefore, it may be tricky to exploit such a relation explicitly/analytically within a sophisticated demix/remix application. However, in the ISS context, basic manipulations such as volume control (up to complete suppression or soloing) or respatialization based on repanning or inversion of the convolutive term, can be implemented since the ISS coder has access to these "generalized source image signals". For example, this can be done by using Wiener filters built from the source image spectrograms, in the same way as what has been done before on uncompressed mix signals [9].

Although basic, these manipulations are of primary importance for many active listening applications, e.g. gaming or music learning applications. Because a simple addition of all the generalized source images $\tilde{s}_{k,j}$ allows the exact recovery of the mixture m_j (up to machine precision), then it can be assumed that a linear remix made with reasonably modified source images will also be of good artistic quality. In particular, the complete muting of a given source for karaoke applications should not affect the quality of the resulting "N-1" mix.

As noted before, in ISS the convolutive term $c_j(n)$ and even the track level processing, can be computed and encoded with the representation of the source image to allow further re-spatialization as in Equation (6).

5.2. Practical implementation

In practice, the distribution of the gain $g_j(n)$ on the source image signals can be done in different ways, within or outside of the DAW. Two ways are presented here, that may involve little change of the production setup in order to allow posterior separation of the source images.

5.2.1. Side-chaining

First, it can be done with the use of side chaining.

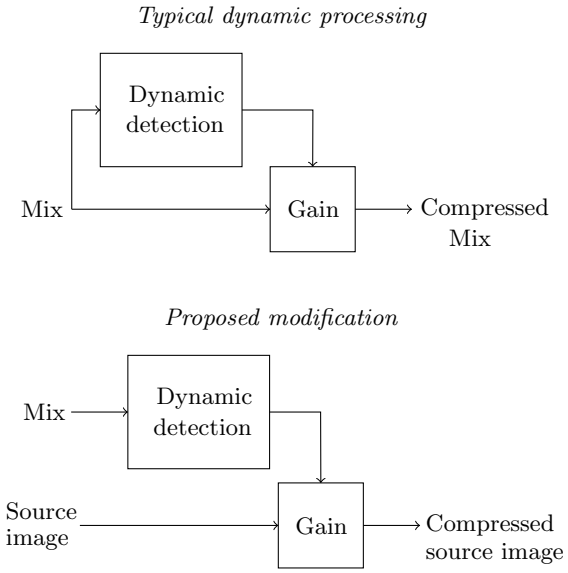

Fig. 3: Dynamic processing. Up: usual implementation; down: use of side chain.

The corresponding configuration of the dynamic processing unit is shown on the bottom of Figure 3. In two passes, the engineer can first record $\hat{m}_j$, which is the mixture without dynamic processing, and then inject it into the dynamic processor side input so that when soloing a set $\mathcal{I}_k$ of tracks, it can still record the corresponding generalized source image with the full effect of the master processors.

Such distribution of the dynamic processing can be done on-line if two mixing busses are used: one containing $\hat{m}_j$ that feeds the side-chain input, and one containing only $\tilde{s}_{k,j}$.

5.2.2. Estimation of the gain reduction

If the mix is already produced, then distribution of the gain reduction may not be available. The remaining option is to pose an inverse problem and to estimate the gain reduction g_j. Then, it would be possible to apply it on the source image signals a posteriori. Therefore, the simplest way to do so is to have the final mix m_j and compare it to the raw mix, i.e. the sum of the pre-mix source image signals $\hat{m}_j(n) = \sum_i \tilde{s}_{k,j}(n)$ (for simplicity of notations, let us consider here the monophonic case of this problem, and omit the channel index j from now on).

Obviously, trying to estimate $g(n)$ by computing

$$\hat{g}(n) = \frac{m(n)}{\hat{m}(n)}$$

would lead to numerical problems when $\hat{m}(n) \to 0$. Amongst the various available possibilities, one can choose to compute time-envelopes using the Hilbert transform $\mathcal{H}$:

$$e(n) = \sqrt{m(n)^2 + \mathcal{H}(m(n))^2}, \quad (8)$$

$$\hat{e}(n) = \sqrt{\hat{m}(n)^2 + \mathcal{H}(\hat{m}(n))^2}. \quad (9)$$

We can estimate $g(n)$ from the envelopes ratio:

$$\hat{g}(n) = \frac{e(n)}{\hat{e}(n)}. \quad (10)$$

Prior smoothing of the envelopes or posterior smoothing of this ratio may be applied to further "regularize" $\hat{g}(n)$, e.g. using a zero phase averaging or median filter. Experimental results are presented on Figure 4. This is a proof of concept on a music mixture of 6 instruments at 44.1kHz sampling rate (Shannon Hurley - Sunrise, Creative Commons). The unprocessed mixture $\hat{m}$ is obtained at level 4 of Figure 2. The mixture $\hat{m}$ is dynamically processed with a professional compressor plugin (Waves RComp) set at 5ms attack, 200ms release, 8:1 compression ratio and a threshold of -10dB. The gain g is estimated using Equation (10) with a 0.5ms median post-filtering. The average signal to prediction error ratio is -37dB.

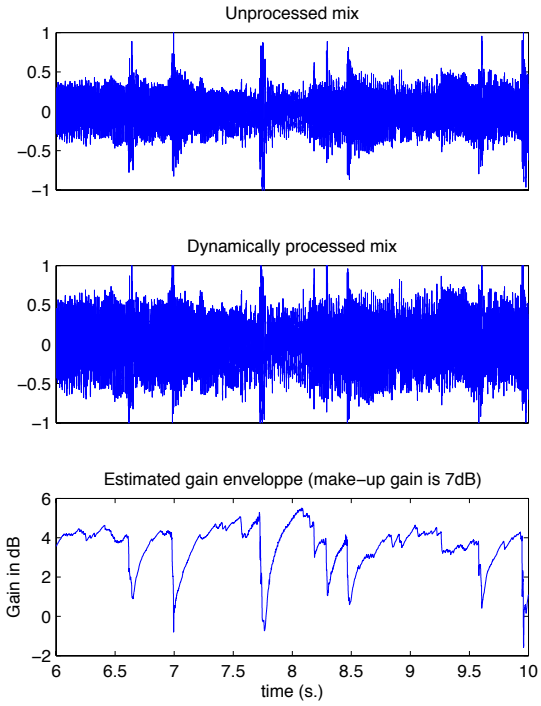

Fig. 4: Estimation of the gain envelope on a mix.

6. CONCLUSION

In this paper we discussed the links and discrepancies between mixing/demixing models in the signal processing literature and the professional music production world. We proposed a "unified" or "generalized" model allowing basic active listening in a linear framework while preserving maximum quality of the artistic mix. This is done by integrating all the linear and most of the non-linear stages of mix processing within the "generalized source image signals": summing these signals leads to exactly recover the artistic mix (up to machine precision).

At the end of the remix chain (i.e. at the general public user level) this technique restitutes the maximum auditory quality while keeping a low complexity, which is a crucial issue for the implementation of active listening systems on mobile platforms, e.g. multimedia players, smartphones or tablets. For instance, such generalized linear framework allows the improvement of the ISS stereo to stereo remixing systems of [9, 4], with no additional complexity at the decoder. As discussed in Section 5, such a system enables basic but important source images manipulation such as volume control and basic respatialization. In the present framework, a musical source can be totally muted without affecting the quality of the resulting music-minus-one mix. At the music production level, the corresponding setup is easily implementable in a classical DAW provided that the dynamic processor on the master track has a side chain input. It can also be implemented a posteriori with little impact on quality, provided that the source image signals before final dynamic processor are available at the active listening encoder.

The tradeoff however, is the increased difficulty at the decoder in accurate respatialization of the so-called generalized source images, that are in fact stereo images already placed in an acoustic space.

Therefore, the proposed model provides a complete separation framework but does not solve the inverse problem of finding back the (ideal) sources composing the mixture. Future work should then focus on a practical implementation of an ISS coding/decoding framework using this model, and on the inversion of the mixing effects present on the estimated signals.

ACKNOWLEDGMENT

This work was supported by the DReaM project (ANR-09-CORD- 006) of the French National Research Agency CONTINT program.

7. REFERENCES

[1] A. S. Bregman. *Auditory scene analysis*. MIT Press: Cambridge, MA, 1990.

[2] S. Disch, C. Ertel, C. Faller, J. Herre, J. Hilpert, A. Hoelzer, P. Kroon, K. Linzmeier, and C. Spenger. Spatial audio coding: Next-generation efficient and compatible coding of multi-channel audio. In *Audio Engineering Society Convention 117*, October 2004.

[3] J. Engdegard, C. Falch, O. Hellmuth, J. Herre, J. Hilpert, A. Hozer, J. Koppens, H. Mundt, H. Oh, H-O; Purnhagen, B. Resch, L. Terentiev, M. L. Valero, and L. Villemoes. MPEG spatial audio object coding, the ISO/MPEG standard for efficient coding of interactive audio scenes. In *Audio Engineering Society Convention 129*, November 2010.

[4] C. Faller, A. Favrot, Y-W Jung, and H-O Oh. Enhancing stereo audio with remix capability. In *Audio Engineering Society Convention 129*, November 2010.

[5] F. Gallot, O. Lagadec, M. Desainte-Catherine, and S. Marchand. iKlax: a new musical audio format for active listening. In *Proc. International Computer Music Conference (ICMC)*, pages 85–88, Belfast, Ireland, 2008.

[6] S. Gorlow and S. Marchand. Informed source separation: Underdetermined source signal recovery from an instantaneous stereo mixture. In *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pages 309–312, New Paltz, NY, USA, October 2011.

[7] J. Herre and L. Terentiv. Parametric coding of audio objects: Technology, performance, and opportunities. In *Audio Engineering Society Conference: 42nd International Conference: Semantic Audio*, July 2011.

[8] C. Jutten and P. Comon. Handbook of blind source separation. Independent component analysis and applications. Academic Press (Elsevier), 2010.

[9] A. Liutkus, J. Pinel, R. Badeau, L. Girin, and G. Richard. Informed source separation through spectrogram coding and data embedding. *Signal Processing*, pending publication.

[10] P. O'Grady, B. A. Pearlmutter, and S. Rickard. Survey of sparse and non-sparse methods in source separation. *International Journal of Imaging Systems and Technology*, 15:18–33, 2005.

[11] A. Ozerov and C. Févotte. Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation. *IEEE Trans. Audio, Speech, Language Process.*, 18(3):550–563, March 2010.

[12] A. Ozerov, A. Liutkus, R. Badeau, and G. Richard. Informed source separation: source coding meets source separation. In *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pages 257—260, New Paltz, NY, USA, October 2011.

[13] M. Parvaix and L. Girin. Informed source separation of linear instantaneous under-determined audio mixtures by source index embedding. *IEEE Trans. Audio, Speech, Language Process.*, 19(6):1721–1733, August 2011.

[14] M. Parvaix, L. Girin, and J.-M. Brossier. A watermarking-based method for informed source separation of audio signals with a single sensor. *IEEE Trans. on Audio, Speech, and Language Processing*, 18(6):1464–1475, 2010.

[15] D. F. Rosenthal and H. G. Okuno. *Computational auditory scene analysis*. Mahwah, NJ: Lawrence Erlbaum, 1998.

[16] A. Taleb and Jutten C. Source separation in post non linear mixtures. *IEEE Trans. on Signal Process.*, 47(10):2807–20, 1999.

[17] E. Vickers. The loudness war: Background, speculation, and recommendations. In *Audio Engineering Society Convention 129*, November 2010.

[18] E. Vincent, R. Gribonval, and C. Févotte. Performance measurement in blind audio source separation. *IEEE Trans. Audio, Speech, Language Process.*, 14(4):1462–1469, July 2006.