Reverse Engineering Stereo Music Recordings Pursuing an Informed Two-Stage Approach

Stanislaw Gorlow, Sylvain Marchand

REVERSE ENGINEERING STEREO MUSIC RECORDINGS PURSUING AN INFORMED TWO-STAGE APPROACH

*Stanislaw Gorlow**
Univ. Bordeaux
LaBRI, UMR 5800
33400 Talence, France
stanislaw.gorlow@labri.fr

Sylvain Marchand
Univ. Brest
Lab-STICC — CNRS, UMR 6285
29238 Brest, France
sylvain.marchand@univ-brest.fr

ABSTRACT

A cascade reverse engineering approach is presented which uses an explicit model of the music production chain. The model considers both the mixing and the mastering stages and incorporates a parametric signal model. The approach is further pursued in an informed scenario. This means that the model parameters are attached in the form of auxiliary data to the mastered mix. They are resorted to afterwards in order to undo the mastering and the mixing. The validity of the approach is demonstrated on a stereo mixture.

1. INTRODUCTION

Most, if not all, professionally produced music recordings undergo two basic processes alias *mixing* and *mastering*. In addition, they are distributed using one of common *stereo* formats such as the Compact Disc Digital Audio. The term "mixing" refers to the process of putting multiple layers of recorded and edited audio together so as to make one final mix, while "mastering" refers to the process of optimizing the mix and transferring it to a storage device.

So far, *informed* source separation techniques consider linear mixing only. Recently [1], efforts have been made to deduce a generalized mixing model that takes the complete music production chain into account. There, it is argued in favor of a linear model which unifies *linear* effects, such as reverberation, with *nonlinear* processing, such as dynamic range compression. However, the motivation for the model is to undo the mixing taking mastering into account but not to undo the mastering as such.

To enable active listening [2], one must reacquire access to latent source components given the mastered mix, i.e. one must *reverse engineer* the mix [3]. In this paper we present a two-stage cascade scheme which models the mixing and the mastering *separately*. Additionally, we demonstrate that knowing the parameter setting that was used for mastering one is able to recover the source components with roughly the same quality as if the mix were not compressed. Also, the separated source signals exhibit high perceptual quality if the mixing parameters and source statistics are known.

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

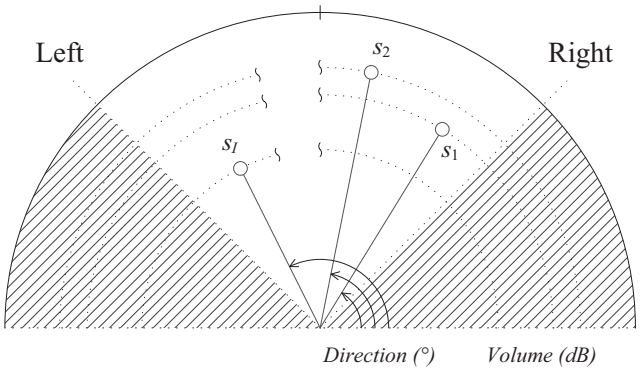

Figure 1: *Modeling of mono sources in a stereo sound field using the parameters direction and volume (implicit).*

As a side note—in [4] a proof is provided which shows that it is not possible to handle a nonlinear mixture without distortion if additional prior knowledge of the nonlinearity is not given, see also [5] and the references therein.

The organization of the paper is as follows. The mixing and the mastering model as well as the problem at hand are given in Section 2. Our approach to reverse engineering the mix is summarized in Section 3. In Section 4, an encoder-decoder framework is shown for practical application. The cascade scheme is evaluated on an exemplary multitrack in Section 5 and then discussed in Section 6. Conclusions are drawn in Section 7.

2. SIGNAL MODEL AND PROBLEM STATEMENT

2.1. Source-Mixture Model (Mixing)

2.1.1. Source Model

We model the signals in the complex subband domain with the short-time Fourier transform (STFT) as a filter bank. A subband signal is modeled as a circular symmetric complex

normal stochastic process with zero mean that evolves over discrete time n. The set of source signal components for a given instant n is deemed to be mutually independent, and so is the set of sources. The sources are hence uncorrelated. A source is mono (single-channel). Each source is assigned a location in the stereo sound field via amplitude panning:

$$\begin{aligned}\mathbf{u}_i(n) &= a_{il}\mathbf{e}_l s_i(n) + a_{ir}\mathbf{e}_r s_i(n) \\ &= \mathbf{a}_i s_i(n),\end{aligned} \tag{1}$$

where $s_i(n)$ is the ith source signal and $\mathbf{a}_i = [\, a_{il} \;\; a_{ir} \,]^\mathsf{T}$ is a time-invariant steering vector. Accordingly, $\mathbf{u}_i$ represents a stereo image of the ith source with $s_i(n) \in \mathbb{C}$ and $\mathbf{a}_i \in \mathbb{R}^2$, where $\{\mathbf{e}_l, \mathbf{e}_r\}$ is the standard basis of $\mathbb{R}^2$. For simplicity, the subband index k is omitted. The ith steering vector $\mathbf{a}_i$ is defined as

$$\mathbf{a}_i \triangleq \begin{bmatrix}\sin\theta_i \\ \cos\theta_i\end{bmatrix}, \tag{2}$$

where θ is the direction parameter, see Fig. 1 [6].

2.1.2. Mixture Model

The mixture is considered to be obtained by superposition of distinct stereo images that were created according to (1). To account for professionally produced music recordings, $s_i(n)$ is regarded as having undergone prior processing in the form of linear and nonlinear audio effects [1]. Thus, the mixture signal is

$$\mathbf{x}_k(n) = \sum_{i\in\mathsf{I}} \underbrace{\mathbf{a}_i s_{ik}(n)}_{\mathbf{u}_{ik}(n)} = \mathbf{A}\mathbf{s}_k(n) \tag{3}$$

with $\mathbf{s}_k = [\, s_{1k} \;\; s_{2k} \;\; \cdots \;\; s_{Ik} \,]^\mathsf{T}$, $\mathbf{A} = [\, \mathbf{a}_1 \;\; \mathbf{a}_2 \;\; \cdots \;\; \mathbf{a}_I \,]$, and $\mathbf{x}_k = [\, x_{lk} \;\; x_{rk} \,]^\mathsf{T}$.

2.2. Compressor Model (Mastering)

2.2.1. Feed-Forward Broadband Compression

Fig. 2 illustrates a basic compressor model [7] which has a switchable RMS/peak detector in the side chain. The input signal $x(n)$ is split and a copy is sent to the side chain. The detector then calculates the sound level or envelope $v(n)$ of $x(n)$ according to

$$\tilde{x}(n) = \beta|x(n)|^p + \bar{\beta}\tilde{x}(n-1) \tag{4a}$$

$$v(n) = \sqrt[p]{\tilde{x}(n)} \tag{4b}$$

where $p = 1$ represents a peak detector and $p = 2$ an RMS detector, respectively. The detector's temporal behavior is controlled by the attack and release parameters through the smoothing factor β, $0 < \beta \leqslant 1$, or $\bar{\beta} = 1 - \beta$. β may take on different values, β_{att} or β_{rel}, depending on whether the detector is in the attack or release phase. The condition for the detector to choose β_{att} over β_{rel} is

$$|x(n)| > v(n-1). \tag{5}$$

A formula that converts a time constant τ into a smoothing factor is given in [8], so e.g.

$$\beta = 1 - \exp\left[-2.2/(f_s \cdot \tau_v)\right], \tag{6}$$

where $\exp$ is the exponential function and f_s the sampling frequency. The sound level $v(n)$ is then compared with the threshold level and, for the case it exceeds the threshold, a scale factor $f(n)$, which corresponds to the ratio of input to output level R, is calculated. This static nonlinearity in the gain computer is modeled in the logarithmic or $\log$ domain as a continuous piecewise linear function:

$$F(n) = \begin{cases} -S \cdot [V(n) - L] & \text{if } V(n) > L, \\ 0 & \text{otherwise,} \end{cases} \tag{7}$$

where L is the threshold in decibel, $V(n) = 20\log_{10} v(n)$, and S is the slope,

$$S = 1 - \frac{1}{R}. \tag{8}$$

The knee parameter determines how quick the compression ratio is reached and shall be ignored, i.e. the knee is "hard". At the end of the side chain the scale factor $f(n)$ is fed to a smoothing filter that yields the gain $g(n)$,

$$g(n) = \gamma f(n) + \bar{\gamma} g(n-1) \qquad \text{with } \gamma \in \{\gamma_{\mathrm{att}}, \gamma_{\mathrm{rel}}\}, \tag{9}$$

where $\bar{\gamma} = 1 - \gamma$ and the decision to choose γ_{att} instead of γ_{rel} is subject to

$$f(n) < g(n-1). \tag{10}$$

The response of the gain smoothing filter is thus controlled by another set of attack and release parameters. Finally, the broadband gain control multiplies the input signal $x(n)$ by the smoothed gain $g(n)$ and adds some makeup gain M to bring the compressed output signal $y(n)$ to a desired level:

$$y(n) = m \cdot [g(n)x(n)], \tag{11}$$

where $m = 10^{M/20}$.

2.2.2. Stereo Linking

In order to avoid image shifting, it is imperative that equal amount of gain reduction be applied to both channels of $\mathbf{x}$. This is achieved by calculating the required amount of gain reduction for $x_l(n)$ and $x_r(n)$ independently, and applying the larger amount to both channels:

$$\mathbf{y}(n) = m \cdot [g(n)\mathbf{x}(n)], \tag{12}$$

where $\mathbf{y} = [\, y_l \;\; y_r \,]^\mathsf{T}$ and

$$g(n) = \min\left[g_l(n), g_r(n)\right]. \tag{13}$$

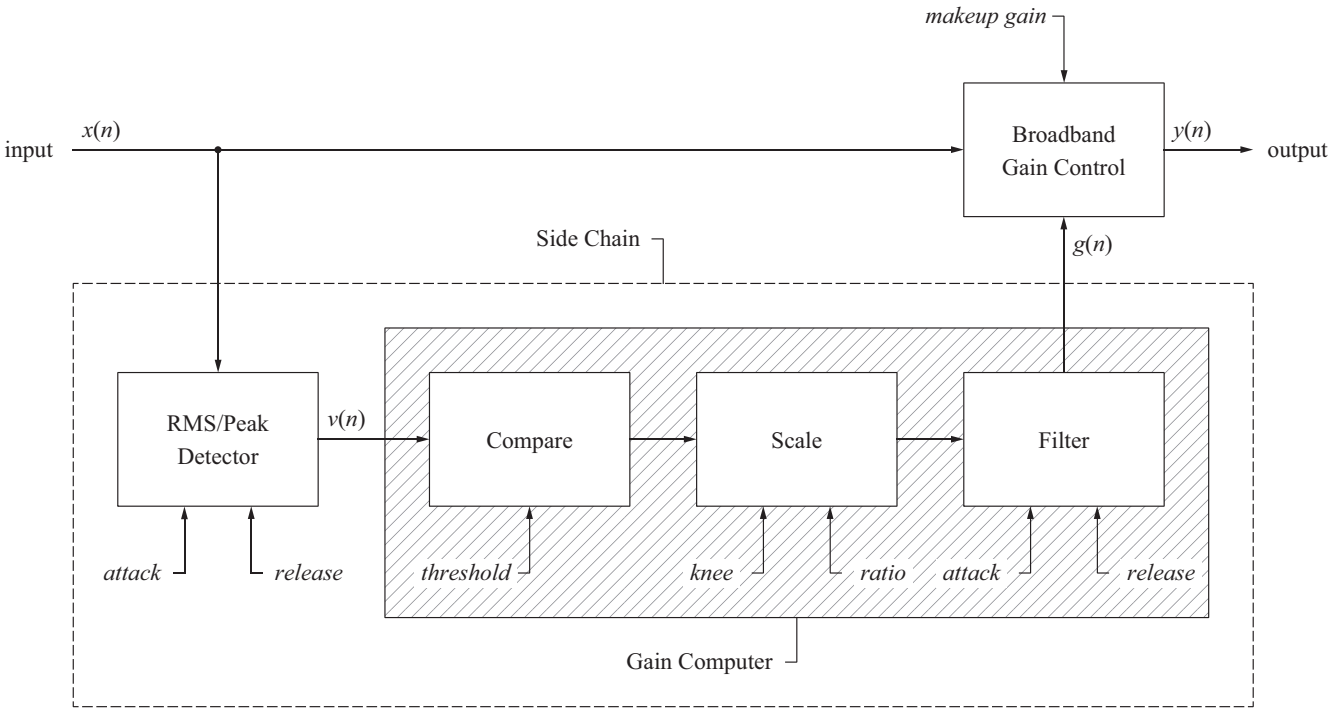

Figure 2: *Basic feed-forward broadband compressor model.*

2.3. Model Parameters

According to the used signal model, the model parameters comprise the following quantities. The direction angles for the sources $\{\theta_i\}$, the instantaneous variance distribution of each source over subbands

$$\phi_{ik}(n) = \mathrm{E}\left[|s_{ik}(n)|^2\right] \qquad \text{for } i = 1, 2, \ldots, I, \qquad (14)$$

where E is the expectation and $\{\phi_{ik}(n)\}_k$ is the short-time power spectral density (STPSD) at instant n, and lastly the compressor parameters

$$\boldsymbol{\psi} = \begin{bmatrix} p & L & R & \beta_{\mathrm{att}} & \beta_{\mathrm{rel}} & \gamma_{\mathrm{att}} & \gamma_{\mathrm{rel}} & M \end{bmatrix}.$$

The STPSDs are estimated according to

$$\hat{\phi}_{ik}(n) = |S_i(k,n)|^2, \qquad (15)$$

where $\{S_i(k,n)\}_k$ represents the frequency spectrum. For a more intuitive use, the compressor parameters β and γ are replaced by the time constants τ_v and τ_g. The meaning of each parameter is recapitulated in Table 1.

2.4. Problem Statement

The problem at hand is stated as follows. Given the mixing, mastering, and signal parameters listed in Table 1, recover the source signals from a mixture signal with a compressed dynamic range in best possible quality. The amount of data associated with the signal parameters shall furthermore be kept to a minimum.

Table 1: *Model parameters and their meaning.*

Parameter	Description
θ	Direction angle in °
ϕ	Power value
p	Detector type (peak or RMS)
L	Threshold level in dB
R	Compression ratio $\mathrm{dB_{in}} : \mathrm{dB_{out}}$
$\tau_{v,\mathrm{att}}$	Attack time of the envelope filter in ms
$\tau_{v,\mathrm{rel}}$	Release time of the envelope filter in ms
$\tau_{g,\mathrm{att}}$	Attack time of the gain filter in ms
$\tau_{g,\mathrm{rel}}$	Release time of the gain filter in ms
M	Makeup gain in dB

3. REVERSE AUDIO ENGINEERING

3.1. Informed Dynamic Range Decompression

In [7] it is shown that, and how, a dynamic nonlinear time-variant operator, such as a dynamic range compressor, can be inverted using an explicit signal model. By knowing the model parameters $\boldsymbol{\psi}$ one is able to recover the original, i.e. uncompressed, signal with high numerical accuracy. What follows is a brief summary of [7].

3.1.1. Characteristic Function

The compressor in Fig. 2 is characterized by the function

$$\begin{aligned}\zeta_p(v) = &\left[\gamma\kappa v^{-S}(n) + \bar{\gamma} g(n-1)\right]^p \\ &\cdot \left[v^p(n) - \bar{\beta}\tilde{x}(n-1)\right] - \beta\left[\frac{|y(n)|}{m}\right]^p,\end{aligned} \tag{16}$$

where $\kappa = l^S$. The root v_0 of $\zeta_p(v)$ bears the instantaneous sound level estimate $\hat{v}(n)$ given the compressed signal $y(n)$ and the compression parameters $\boldsymbol{\psi}$. All other unknowns are computed from $v_0(n)$ according to

$$\begin{aligned}\tilde{x}(n) &= v_0^p(n) \\ |\hat{x}(n)| &= \sqrt[p]{\left[\tilde{x}(n) - \bar{\beta}\tilde{x}(n-1)\right]/\beta} \\ g(n) &= |y(n)|/[m \cdot |\hat{x}(n)|]\end{aligned} \tag{17}$$

if $\hat{v}(n) > 10^{L/20}$, i.e. the current sample is compressed, or

$$\begin{aligned}g(n) &= \gamma + \bar{\gamma} g(n-1) \\ |\hat{x}(n)| &= |y(n)|/[m \cdot g(n)] \\ \tilde{x}(n) &= \beta|\hat{x}(n)|^p + \bar{\beta}\tilde{x}(n-1)\end{aligned} \tag{18}$$

otherwise. The decompressed sample is then given by

$$\hat{x}(n) = \mathrm{sgn}(y) \cdot |\hat{x}(n)|, \tag{19}$$

where sgn is the signum function.

3.1.2. Attack-Release Phase Toggle

When a peak detector is in use, β can take on two different values. The condition for the attack phase is

$$\left[\frac{|y(n)|}{m \cdot g(n-1)}\right]^p > \tilde{x}(n-1). \tag{20}$$

Likewise, the gain smoothing filter can be in the attack or release phase. The following condition is used to detect the attack phase of the gain smoothing filter:

$$\begin{aligned}&\sqrt[p]{\beta\left[\frac{|y(n)|}{m \cdot g(n-1)}\right]^p + \bar{\beta}\tilde{x}(n-1)} \\ &\qquad > \left[\frac{\kappa}{g(n-1)}\right]^{1/s}.\end{aligned} \tag{21}$$

3.1.3. Envelope Predictor

An estimate of the envelope value $\hat{v}(n)$ is needed to detect when compression is active, formally $V(n) > L$ in (7). The corresponding equation is

$$\hat{v}(n) = \sqrt[p]{\beta\left[\frac{|y(n)|}{m \cdot [\gamma + \bar{\gamma} g(n-1)]}\right]^p + \bar{\beta}\tilde{x}(n-1)}, \tag{22}$$

where β and γ are selected using (20) and (21).

3.1.4. Stereo Unlinking

First, one decompresses both the left and the right channel of $\mathbf{y}$ independently using $\boldsymbol{\psi}$, and one obtains two estimates $\hat{x}_l(n)$ and $\hat{x}_r(n)$. Using (11), one then computes $\hat{y}_l(n)$ and $\hat{y}_r(n)$ from $\hat{x}_l(n)$ and $\hat{x}_r(n)$, and picks the channel ref for which $\hat{y}_{\mathrm{ref}}(n) \approx y_{\mathrm{ref}}(n)$. Finally, one updates the variables of the complementary channel $\neg\mathrm{ref}$:

$$\hat{x}_{\neg\mathrm{ref}}(n) = \frac{y_{\neg\mathrm{ref}}(n)}{m \cdot g_{\mathrm{ref}}(n)}, \tag{23}$$

$\tilde{x}_{\neg\mathrm{ref}}(n)$ according to (4a), and $g_{\neg\mathrm{ref}}(n)$ according to (9).

3.2. Informed Audio Source Separation

In [6] we discuss how an underdetermined stereo mixture is decomposed into distinct source signal components using a constrained spatial filtering approach. The approach, which presumes that the model parameters $\{\theta_i\}$ and $\{\phi_{ik}(n)\}$ are known, is summarized below.

3.2.1. Spatial Covariance Matrix

The local mixture spatial covariance matrix is given by

$$\begin{aligned}\mathbf{R}_{\mathbf{xx},k}(n) &= \mathrm{E}\left[\mathbf{x}_k(n)\mathbf{x}_k^{\mathsf{H}}(n)\right] \\ &= \sum_{i\in\mathsf{I}} \mathbf{a}_i\mathbf{a}_i^{\mathsf{T}}\phi_{ik}(n),\end{aligned} \tag{24}$$

where H denotes Hermitian transpose. Using (24) and (2), the mixture spatial covariance matrix is reconstructed from the direction angles $\{\theta_i\}$ and the STPSDs $\{\phi_{ik}(n)\}_k$, $i = 1, 2, \ldots, I$. The eigenvectors of $\mathbf{R}_{\mathbf{xx}}$ indicate the angles of the maximum and the minimum mean sound power that is encountered in the time-frequency (TF) point (k, n).

3.2.2. Power-Conserving Minimum-Variance Filter

When more than two sources are active in a time-frequency point (k, n), the ith signal component is separated from the mixture with the help of the "power-conserving minimum-variance" (PCMV) spatial filter [6] $\hat{\mathbf{w}}_{ik}(n)$,

$$\hat{\mathbf{w}}_{ik}(n) \triangleq \mathbf{R}_{\mathbf{xx},k}^{-1}(n)\mathbf{a}_i\sqrt{\frac{\phi_{ik}(n)}{\mathbf{a}_i^{\mathsf{T}}\mathbf{R}_{\mathbf{xx},k}^{-1}(n)\mathbf{a}_i}}, \tag{25}$$

$i = 1, 2, \ldots, I_k(n)$, according to

$$\hat{s}_{ik}(n) = \hat{\mathbf{w}}_{ik}^{\mathsf{T}}(n)\mathbf{x}_k(n). \tag{26}$$

If the number of active sources is at most two, the demixing is trivial, given that the mixing system $\mathbf{A}$ is known. In that case, the separation reduces to the inversion of $\mathbf{A}$.

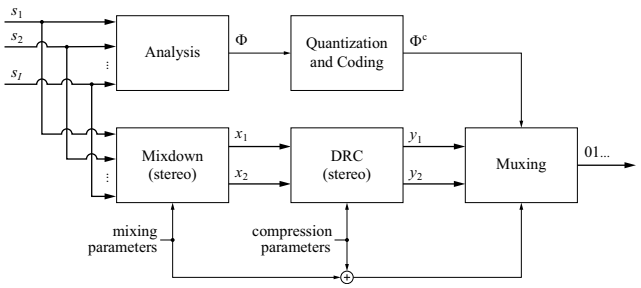

Figure 3: *A two-stage cascade encoder.*

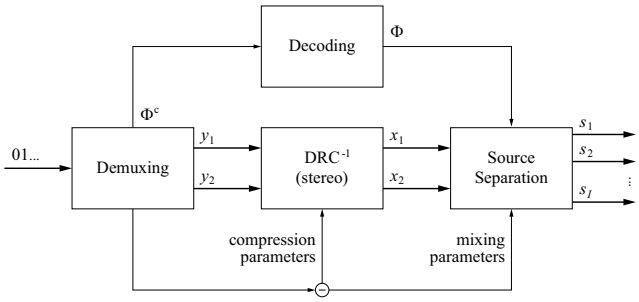

Figure 4: *A two-stage cascade decoder.*

4. PROPOSED CASCADE SCHEME

To account for a simplified yet complete music production chain that consists of mixing and mastering, we propose to combine the two processing steps from Section 3 in a two-stage cascade decoder scheme. The corresponding encoder and decoder block diagrams are illustrated in Figs. 3–4.

4.1. Encoder

Fig. 3 shows the cascade scheme for a practical encoder. It contains an analysis block that computes the STPSDs $\mathbf{\Phi} = [\,\boldsymbol{\phi}_1\ \boldsymbol{\phi}_2\ \cdots\ \boldsymbol{\phi}_I\,]^{\mathsf{T}}$ with $\boldsymbol{\phi}_i = [\,\phi_{i1}\ \phi_{i2}\ \cdots\ \phi_{iK}\,]$, where K is the number of frequency bands, from the source signals $\mathbf{s}_k(n)$, a quantization and coding block that reduces the size of the metadata, a mixdown block that represents the mixing, and a dynamic range compression (DRC) block that represents the mastering stage. The multiplexing block assembles the bitstream, which includes the compressed stereo signal, the coded metadata, and other model parameters.

4.2. Decoder

The decoder is shown in Fig. 4. First, the demuxing block disassembles the bitstream into the compressed signal, the metadata, and other model parameters. The DRC^{-1} block provides the decompressed mixture signal $\hat{\mathbf{x}}(n)$. The latter is then decomposed in the source separation block with the aid of the metadata $\mathbf{\Phi}$, yielding the source signal estimates $\hat{\mathbf{s}}(n) = [\,\hat{s}_1(n)\ \hat{s}_2(n)\ \cdots\ \hat{s}_I(n)\,]^{\mathsf{T}}$ in the original time domain.

5. PERFORMANCE EVALUATION

5.1. Performance Metrics

For the purpose of evaluation of the proposed scheme, the following two metrics are used: the root-mean-square error (RMSE) defined as

$$\mathrm{RMSE}_i = \sqrt{\frac{1}{|\mathsf{N}|}\sum_{n\in\mathsf{N}}\left[\hat{s}_i(n) - s_i(n)\right]^2}, \qquad (27)$$

Table 2: *Compressor setting used for the complete mix.*

Parameter	Description	Value
p	Type	RMS
L (dBFS)	Threshold	-32.0
R ($\mathrm{dB_{in}} : \mathrm{dB_{out}}$)	Ratio	$3.0 : 1$
$\tau_{v,\mathrm{att}}$ (ms)	Envelope attack	5.0
$\tau_{v,\mathrm{rel}}$ (ms)	Envelope release	
$\tau_{g,\mathrm{att}}$ (ms)	Gain attack	13.0
$\tau_{g,\mathrm{rel}}$ (ms)	Gain release	435
M (dB)	Makeup	9.0

where $s_i(n)$ represents a time-domain signal, and PSM as an objective measure for perceptual similarity between the original signal $s_i(n)$ and its estimate $\hat{s}_i(n)$. The RMSE is given in dB relative to full scale (dBFS). PSM is computed with PEMO-Q [9, 10]. In [9] it is said that PEMO-Q shows a slightly better performance than Perceptual Evaluation of Audio Quality (PEAQ) [11].

5.2. Experimental Design

We use the algorithm from [7] for decompression together with the source separation framework from [6]. We employ a 2048-point fast Fourier transform together with a Kaiser–Bessel derived window of the same size and let succeeding frames overlap by 50 %. The two-stage scheme is tested on Fort Minor's "Remember the Name" multitrack, which has been decomposed into 5 mono sources and cut down to 24 s in length. The compressor setting is listed in Table 2. To exclude a performance bias due to quantization, the values in Table 2 and the sources' locations are considered known on the decoder side. Moreover, since the sources' locations and the compressor setting are time-invariant, their size can be neglected. The STPSD $\boldsymbol{\phi}_i$ is uniformly quantized with 6 bits on a 76-band nonuniform frequency scale. By applying Huffman coding to differentially pulse-code modulated ϕ-values, the mean metadata rate reduces to roughly 10 kbps per source. The simulations are run in MATLAB.

Table 3: *RMSE and SNR for the three mixture signals.*

Mixture type	RMSE (dBFS)	SNR (dB)
Compressed	-31.8	3.08
Compressed*	-36.0	7.27
Decompressed	-62.3	33.6

5.3. Experimental Results

The results are depicted in Fig. 5. The accompanying audio can be found on http://www.labri.fr/~gorlow/dafx13/. The asterisk marks the compressed mix *without* makeup gain, i. e. $M = 0$ dB. The RMSE and signal-to-noise ratio (SNR) for each mixture signal,

$$\mathrm{SNR} \triangleq 10 \log_{10} \frac{\sum_j \|\mathbf{x}_j\|^2}{\sum_j \|\mathbf{y}_j - \mathbf{x}_j\|^2}, \tag{28}$$

where $\|\cdot\|$ is the Euclidean norm, are given in Table 3.

It can be noted that the separated source signals exhibit relatively high quality when the mix is uncompressed (see dashed bar). The RMSE for the vocal track reaches almost -60 dBFS, while being below -40 dBFS for the rest. The decompressor's performance is practically free from error: the $\Delta\mathrm{RMSE}$ and $\Delta\mathrm{PSM}$ values for each track are close to zero after decompression (see lower row). The RMSE level decreases, as expected, if the makeup gain is removed from the compressed mix, but does not reach the level of the uncompressed mix. This proves that the waveform of the mix has been altered by the compressor. The RMSE difference between the two compressed mixtures so is due to scaling. On the contrary, the corresponding PSM values are equal, which shows that the PSM metric is scale-independent.

In the given example, the PSM improvement due to the decompressor is mostly evident for the bass track. For the other tracks, the PCMV filter provides an estimate which is perceptually very similar to the reference, even if the mix is compressed. However, a so-called "pumping" coming from hard compression can be heard clearly on the vocal track. The effect is more audible for faster attack and release. So, for the used compressor setting, to achieve high perceptual quality the decompressor is indispensable.

6. DISCUSSION

The proposed scheme is based on a simplified model of the music production chain, which consists of a summation of amplitude-panned single-channel recordings in the mixing stage, followed by broadband dynamic range compression in the mastering stage. Although one could certainly argue that commercial releases are created from both single- *and* two-channel tracks and that *more sophisticated* compressor models are employed in electronic music or *no compressor* at all as in classical or acoustic music, albeit "transparent" compression may have still been applied, the cascade shall be viewed as a "blueprint" rather than a "standard". As an example, compression can be avoided easily by setting the threshold to 0 dBFS or by "bypassing" the compressor and the decompressor. In regard to electronic and also pop/rock music, it should be possible to determine the characteristic function of the respective compressor and to solve it using the approach in [7]. Besides, the spatial filtering approach is likewise applicable to two-channel tracks [12]. If a time difference between the left and the right channel is wished for and the delay in samples is sufficiently small compared to the STFT length, so that the two windowed signals carry almost the same content, a frequency-dependent phase shift can be added to the steering vector. The new vector $\mathbf{a}_{ik}$ is then complex and so will be the spatial filter $\hat{\mathbf{w}}_{ik}(n)$.

The upper performance bound of the scheme is mainly due the sound complexity of commercial music. Notably, it depends on by how much the frequency spectra of distinct tracks overlap and how the sources are distributed in space. The sound quality after demixing is subject to the so-called "array gain", which can be shown to be a function of a) the STPSDs and b) the mixing system; see [13]. As a general rule, the less the source signals' frequency spectra overlap and the further apart they are placed the better the resulting sound quality. Yet both constraints are rarely met in reality. Tracks are either in the same key, or their keys are relative or in a subdominant or dominant relationship. So, stringed instruments have a high number of interfering harmonics in the mix. Percussion instruments, on the other hand, cover a broad range of frequencies. Furthermore, traditional source positioning is such that the main sound sources are near the center, i.e. very close to each other. When multiple sources are in one direction, they can only be distinguished by their spectral envelope, which diminishes quality as well.

The knowledge of the mixing and the mastering can be viewed as a shortcoming of the scheme, rendering it hardly applicable to existent music releases, for which that sort of information is unavailable. And although tools, such as the non-negative matrix factorization (NMF) [14, 15], to learn the spectrograms of the source signals from the mixture do exist, their performance is often limited. First, because the factorization is not unique, and second, the cost function is non-convex. In our scheme, like in any other model-based scheme, a deviation from the true parameters will cause an additional error in the result. The decoder is most effective when supplied with sufficiently accurate information by an accompanying encoder.

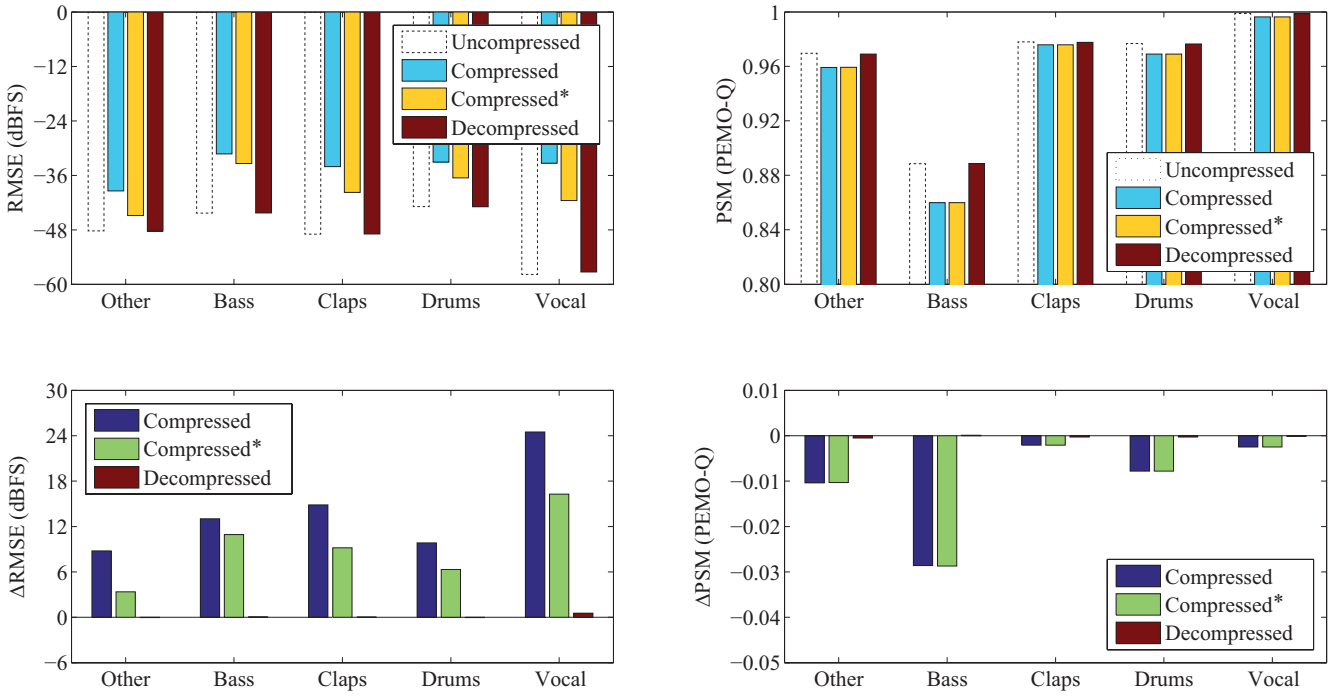

Figure 5: *RMSE and PSM values for the multitrack (upper row) and the corresponding difference values between the estimates from the de-/compressed and the uncompressed mixture signal (lower row). The asterisk ($*$) indicates $M = 0$ (no gain).*

7. CONCLUSION AND FUTURE WORK

We demonstrated a two-stage reverse engineering approach consisting of decompression and source separation. By an example it was shown that if mastering can be undone with a negligibly small error, the demixed sources show a sound quality almost identical with the case where the mix is not compressed. The decompressor's numerical accuracy could so be pivotal in more complex schemes which include, e.g., deconvolution in the demixing stage. This approach is also rate-efficient, as it only requires that the model parameters are known on the decoder side.

Future work could focus on the effects that "lossy" data compression, such as MP3 or AAC [16], and watermarking [17, 18] have on the system's performance. In addition, the cascade should be tested for different compressor types and settings and on a larger dataset.

8. REFERENCES

[1] N. Sturmel *et al.*, "Linear mixing models for active listening of music productions in realistic studio conditions," in *AES Conv. 132*, Apr. 2012, pp. 1–10.

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

[3] D. Barchiesi and J. Reiss, "Reverse engineering of a mix," *J. Audio Eng. Soc.*, vol. 58, no. 7/8, pp. 563–576, Jul. 2010.

[4] A. Taleb and C. Jutten, "Source separation in post-nonlinear mixtures," *IEEE Trans. Signal Process.*, vol. 47, no. 10, pp. 2807–2820, Oct. 1999.

[5] C. Jutten and J. Karhunen, "Advances in nonlinear blind source separation," in *Proc. ICA*, Apr. 2003, pp. 245–256.

[6] S. Gorlow and S. Marchand, "Informed audio source separation using linearly constrained spatial filters," *IEEE Audio, Speech, Language Process.*, vol. 21, no. 1, pp. 3–13, Jan. 2013.

[7] S. Gorlow and J. D. Reiss, "Model-based inversion of dynamic range compression," *IEEE Audio, Speech, Language Process.*, vol. 21, no. 7, pp. 1434–1444, Jul. 2013.

[8] U. Zölzer, *DAFX: Digital audio effects*, 2nd ed. John Wiley & Sons Ltd, 2011, ch. 4.

[9] R. Huber and B. Kollmeier, "PEMO-Q — a new method for objective audio quality assessment using a model of auditory perception," *IEEE Audio, Speech, Language Process.*, vol. 14, no. 6, pp. 1902–1911, Nov. 2006.

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

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

[12] S. Gorlow and S. Marchand, "Informed separation of spatial images of stereo music recordings using second-order statistics," 2013, manuscript submitted for publication.

[13] S. Gorlow, E. A. P. Habets, and S. Marchand, "Multichannel object-based audio coding with controllable quality," in *Proc. IEEE ICASSP*, May 2013, pp. 561–565.

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

[15] ——, "Algorithms for non-negative matrix factorization," in *Proc. NIPS*, Dec. 2000, pp. 556–562.

[16] K. Brandenburg, "MP3 and AAC explained," in *AES 17th International Conference on High-Quality Audio Coding*, Aug. 1999, pp. 1–12.

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

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