HAL
open science

Multichannel audio declipping

Alexey Ozerov, Cagdas Bilen, Patrick Pérez

To cite this version:

Alexey Ozerov, Cagdas Bilen, Patrick Pérez. Multichannel audio declipping. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'16), Mar 2016, Shanghai, China. hal-01254950v2

HAL Id: hal-01254950

https://inria.hal.science/hal-01254950v2

Submitted on 14 Jan 2016

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

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

MULTICHANNEL AUDIO DECLIPPING

Alexey Ozerov, Çağdaş Bilen and Patrick Pérez

Technicolor
975 avenue des Champs Blancs, CS 17616, 35576 Cesson Sévigné, France
{alexey.ozerov, cagdas.bilen, patrick.perez}@technicolor.com

ABSTRACT

Audio declipping consists in recovering so-called clipped audio samples that are set to a maximum / minimum threshold. Many different approaches were proposed to solve this problem in case of single-channel (mono) recordings. However, while most of audio recordings are multichannel nowadays, there is no method designed specifically for multichannel audio declipping, where the inter-channel correlations may be efficiently exploited for a better declipping result. In this work we propose for the first time such a multichannel audio declipping method. Our method is based on representing a multichannel audio recording as a convolutive mixture of several audio sources, and on modeling the source power spectrograms and mixing filters by nonnegative tensor factorization model and full-rank covariance matrices, respectively. A generalized expectation-maximization algorithm is proposed to estimate model parameters. It is shown experimentally that the proposed multichannel audio declipping algorithm outperforms in average and in most cases a state-of-the-art single-channel declipping algorithm applied to each channel independently.

***Index Terms*—** Multichannel audio, audio declipping, full-rank spatial model, nonnegative tensor factorization, generalized expectation-maximization

1. INTRODUCTION

Audio declipping consists in recovering so-called *clipped* audio samples that are set to a maximum / minimum threshold. This signal clipping may happen due to limits of the acquisition system or during audio post-processing. Audio declipping belongs as well to a larger family of *audio inpainting* problems, as recently formulated by Adler *et al.* [1], where the goal is to recover any missing audio samples, assuming their locations within the observed signal are known. Audio declipping is a well known problem and several approaches were proposed in the past [2–4]. Moreover, since the publication of [1], where audio inpainting is posed as a general inverse problem, audio declipping has regained interest and several new improved methods were proposed [5–9]. Some of these methods are based on local sparse/cosparse models [1, 5, 8] and others on more structural models such as social sparsity [6] or nonnegative matrix factorization (NMF) [9].

All the audio declipping approaches mentioned above [1–9] are designed for single-channel (mono) audio signals. However, the majority of audio recordings nowadays are multichannel, typically stereo. This concerns not only professionally produced audio recordings, but also the user-generated audio captures, since more and more devices (e.g., cameras, smart-phones and tablets) are provided with several microphones. A straightforward and naive approach to declip multichannel audio would consist in applying a single-channel declipping algorithm to each channel independently. However, such a simple strategy would suffer from the following limitations. First, even though the same algorithm is applied to each channel, this is done independently and the reconstruction errors might be uncorrelated over channels, thus leading to spatial inconsistency in the resulting declipped multichannel audio. Second, this approach does not exploit possible correlations between channels, which are usually strong, and thus leads to a suboptimal performance. Both issues could be potentially addressed by an approach modeling and exploiting these correlations. Let us give a simple example. An audio source within a stereo recording can be clipped only in the left channel and not in the right one. As such, it seems very beneficial to use the right channel in order to reconstruct the clipped parts in the left one. This is somehow in line with approaches proposed for image desaturation (an equivalent problem in image processing) [10, 11], where a non-clipped color channel is used to reconstruct a clipped one.

To the best of the authors knowledge none of the existing audio declipping approaches is suitably designed for multichannel audio, i.e., none of them exploits the inter-channel correlation for an improved and spatially consistent declipping result. In this paper we propose such an approach for the first time. We model the multichannel audio to be declipped as a convolutive mixture of several audio sources, which is a usual assumption in audio source separation [12, 13]. We model then (in the short-time Fourier transform [STFT] domain) the source power spectra with an Itakura-Saito (IS) nonnegative tensor factorization (NTF) model [14, 15] and their convolutive mixing with full-rank covariance matrices [16]. This is precisely the modeling used for informed source separation in [17], though here we target a completely different application. The overall modeling leads to a multivariate Gaussian distribution of the mixture STFT, which is a linear transform. Both the Gaussian modeling and the transform's linearity make possible handling time domain losses (due to the clipping) in an optimal way. The proposed approach is essentially an extension of our previous work on single-channel audio declipping using single-source NMF model [9] and on joint single-channel audio declipping and source separation using multi-source NTF model [18]. Note that in the single-channel case, in order to perform audio declipping only, one does not need caring about a possible multi-source nature of an audio recording [1–9], and a multi-source assumption may be needed when the source separation is targeted as well [18]. However, in the multichannel case we consider here, a multi-source assumption becomes essential even for declipping only, since different sources may be clipped differently in different channels due to their different mixing characteristics. As a byproduct of this multi-source assumption, the proposed approach

This work was partially supported by ANR JCJC program MAD (ANR-14-CE27-0002).

may be used without any modification for joint multichannel audio declipping and source separation. However, we do not investigate this feature here and leave it for a further study.

The rest of the paper is organized as follows. The proposed approach, including modeling, model estimation via a generalized expectation-maximization (GEM) algorithm [19], and signal reconstruction via Wiener filtering, is presented in Section 2. Section 3 is devoted to experiments, where the proposed multichannel declipping algorithm is compared with an NMF-based single-channel declipping algorithm [9] applied independently to each channel. In Section 4 conclusions are drawn and some further research directions are given.

2. PROPOSED APPROACH

2.1. Initial assumptions

Let us consider a multichannel audio signal x''_{it}, where $i = 1, \ldots, I$ and $t = 1, \ldots, T$ are channel and time indices, respectively.[1] It is assumed that the signal is clipped everywhere except on so-called *observation support (OS)* $\Xi'' \subset \{1, \ldots, I\} \times \{1, \ldots, T\}$. In other words, the values of x''_{it} are observed for $(i, t) \in \Xi''$ and are clipped and unknown for $(i, t) \in \overline{\Xi}'' \triangleq \{1, \ldots, I\} \times \{1, \ldots, T\} \backslash \Xi''$. The goal of audio declipping is to recover the clipped samples x''_{it}, $(i, t) \in \overline{\Xi}''$, given the observed ones.

2.2. Model

It is assumed that the multichannel audio signal is a mixture of J audio sources as

$$x''_{it} = \sum\nolimits_{j=1}^{J} y''_{ijt}, \tag{1}$$

where y''_{ijt} are the samples of the *source images*, i.e., of the contributions of sources into the mixture, and $j = 1, \ldots, J$ is the source index. It is important to highlight that, while in audio source separation [12, 13] the mixing equation (1) is always introduced within the problem formulation, here it is a part of the modeling assumptions.

In the windowed time domain, the time domain signals are taken in overlapping (usually non-rectangular) frames of length M. Mixture and source images become $\{x'_{imn}\}$ and $\{y'_{ijmn}\}$ respectively, with $n = 1, \ldots, N$ being the frame index and $m = 1, \ldots, M$ the sample location within the frame. The OS within the framed representation corresponding to Ξ'' in time domain is a set $\Xi' \subset \{1, \ldots, I\} \times \{1, \ldots, M\} \times \{1, \ldots, N\}$ whose restrictions to n-th frame and to a couple of i-th channel and n-th frame read $\Xi'_n = \{(i, m) | (i, m, n) \in \Xi'\}$ and $\Xi'_{in} = \{m | (i, m, n) \in \Xi'\}$, respectively. The Short-Time Fourier Transfrom (STFT) coefficients of the time domain signals x''_{it} and y''_{ijt} can be obtained by applying the complex-valued unitary Discrete Fourier Transform (DFT) matrix, $\mathbf{U} \in \mathbb{C}^{F \times M}$, to the windowed time domain counterparts, yielding $\mathbf{x}_{in} = \mathbf{U}\mathbf{x}'_{in}$ and $\mathbf{y}_{ijn} = \mathbf{U}\mathbf{y}'_{ijn}$ where each vector represents the samples within a frame, for example $\mathbf{x}'_{in} = [x'_{imn}]_{m=1\cdots M}$ and $\mathbf{x}_{in} = [x_{ifn}]_{f=1\cdots F}$, with index f representing the frequency index of the Fourier transform coefficients.

[1] Throughout this paper the time-domain signals will be represented by letters with two primes, e.g., x'', framed and windowed time-domain signals will be denoted by letters with one prime, e.g., x', and complex-valued short-time Fourier transform (STFT) coefficients will be denoted by letters with no primes, e.g., x.

With the above assumptions the time domain mixing equation (1) can be rewritten in the STFT domain as

$$\mathbf{x}_{fn} = \sum\nolimits_{j=1}^{J} \mathbf{y}_{jfn}, \tag{2}$$

where $\mathbf{x}_{fn}$ and $\mathbf{y}_{jfn}$ are channel-wise vectors defined as $\mathbf{x}_{fn} = [x_{ifn}]_{i=1\cdots I}$ and $\mathbf{y}_{jfn} = [y_{ijfn}]_{i=1\cdots I}$.

It is then assumed that the latent source images are modeled as in [17] with an IS-NTF spectral model [14, 15] and a full-rank covariance spatial model [16]. More precisely, each complex-valued vector $\mathbf{y}_{jfn}$ is assumed following a zero-mean circular complex Gaussian distribution as

$$\mathbf{y}_{jfn} \sim \mathcal{N}_c(0, \mathbf{R}_{jf} v_{jfn}), \tag{3}$$

where $\mathbf{R}_{jf}$ is a complex-valued positive definite Hermitian matrix (the full-rank spatial model) and $v_{jfn} > 0$ is a variance having the following low-rank NTF structure

$$v_{jfn} = \sum\nolimits_{k=1}^{K} q_{jk} w_{fk} h_{nk}, \tag{4}$$

with $\mathbf{Q} = [q_{jk}]_{j,k}$, $\mathbf{W} = [w_{fk}]_{f,k}$ and $\mathbf{H} = [h_{nk}]_{n,k}$ being, respectively, $J \times K$, $F \times K$ and $N \times K$ nonnegative matrices. The full set of model parameters to be estimated is $\boldsymbol{\theta} = \{\mathbf{Q}, \mathbf{W}, \mathbf{H}, [\mathbf{R}_{jf}]_{j,f}\}$.

2.3. Model estimation and signal reconstruction

Let us first introduce few notations. Let $\bar{\mathbf{x}}'_{in} \triangleq [x'_{imn}]_{m \in \Xi'_{in}}$ denote the observed (non-clipped) windowed samples in the i-th channel and the n-th frame. Let $\bar{\mathbf{x}}'_n \triangleq [\bar{\mathbf{x}}'^T_{1n}, \bar{\mathbf{x}}'^T_{2n}, \ldots, \bar{\mathbf{x}}'^T_{In}]^T$ denote the concatenation over all channels of the observed samples in the n-th frame.

2.3.1. *Estimation of the signal*

One can write the posterior distribution of each source image time-frequency vector $\mathbf{y}_{jfn}$ given the corresponding observed frame $\bar{\mathbf{x}}'_n$ and model $\boldsymbol{\theta}$ as $\mathbf{y}_{jfn} | \bar{\mathbf{x}}'_n; \boldsymbol{\theta} \sim \mathcal{N}_c(\hat{\mathbf{y}}_{jfn}, \widehat{\boldsymbol{\Sigma}}_{\mathbf{y}_{jfn}\mathbf{y}_{jfn}})$ with $\hat{\mathbf{y}}_{jfn}$ and $\widehat{\boldsymbol{\Sigma}}_{\mathbf{y}_{jfn}\mathbf{y}_{jfn}}$ being, respectively, posterior mean and posterior covariance matrix, each of which can be computed by Wiener filtering [20] as[2]

$$\hat{\mathbf{y}}_{jfn} = \boldsymbol{\Sigma}^H_{\bar{\mathbf{x}}'_n \mathbf{y}_{jfn}} \boldsymbol{\Sigma}^{-1}_{\bar{\mathbf{x}}'_n \bar{\mathbf{x}}'_n} \bar{\mathbf{x}}'_n, \tag{5}$$

$$\widehat{\boldsymbol{\Sigma}}_{\mathbf{y}_{jfn}\mathbf{y}_{jfn}} = \boldsymbol{\Sigma}_{\mathbf{y}_{jfn}\mathbf{y}_{jfn}} - \boldsymbol{\Sigma}^H_{\bar{\mathbf{x}}'_n \mathbf{y}_{jfn}} \boldsymbol{\Sigma}^{-1}_{\bar{\mathbf{x}}'_n \bar{\mathbf{x}}'_n} \boldsymbol{\Sigma}_{\bar{\mathbf{x}}'_n \mathbf{y}_{jfn}}, \tag{6}$$

given the definitions

$$\boldsymbol{\Sigma}_{\mathbf{y}_{jfn}\mathbf{y}_{jfn}} \triangleq \mathbf{R}_{jf} v_{jfn}, \tag{7}$$

$$\boldsymbol{\Sigma}_{\bar{\mathbf{x}}'_n \mathbf{y}_{jfn}} \triangleq \widetilde{\mathbf{U}}(\Xi'_n)^H \boldsymbol{\Sigma}_{\mathbf{y}_{jn}\mathbf{y}_{jfn}}, \tag{8}$$

$$\boldsymbol{\Sigma}_{\bar{\mathbf{x}}'_n \bar{\mathbf{x}}'_n} \triangleq \widetilde{\mathbf{U}}(\Xi'_n)^H \sum\nolimits_j \boldsymbol{\Sigma}_{\mathbf{y}_{jn}\mathbf{y}_{jn}} \widetilde{\mathbf{U}}(\Xi'_n), \tag{9}$$

where

- $\boldsymbol{\Sigma}_{\mathbf{y}_{jn}\mathbf{y}_{jn}} \triangleq \left[\mathrm{diag}\left([\mathbf{R}_{jf}(k,l) v_{jfn}]_f\right)_{k,l}\right]$,
- $\boldsymbol{\Sigma}_{\mathbf{y}_{jn}\mathbf{y}_{jfn}}$ is an $IF \times I$ matrix formed by columns of $\boldsymbol{\Sigma}_{\mathbf{y}_{jn}\mathbf{y}_{jn}}$ with index in $\{f, f+F, f+2F, \ldots, f+(I-1)F\}$,
- $\widetilde{\mathbf{U}}(\Xi'_n) \triangleq \mathrm{diag}\left([\mathbf{U}(\Xi'_{in})]_i\right)$ is an $IF \times I|\Xi'_n|$ matrix[3], and

[2] $\mathbf{a}^H$ represents the conjugate transpose of the vector (or matrix) $\mathbf{a}$.

[3] $\mathrm{diag}\left([\mathbf{B}_i]_i\right)$ with $\mathbf{B}_i$ $(i = 1, \ldots, I)$ being matrices represents the corresponding block diagonal matrix.

- $\mathbf{U}(\Xi'_{in})$ is the $F \times |\Xi'_{in}|$ matrix formed by columns from $\mathbf{U}$ with index in Ξ'_{in}.

An estimate of declipped multichannel signal in the STFT domain can be reconstructed by summing up source image estimates computed in (5), i.e., $\hat{\mathbf{x}}_{fn} = \sum_j \hat{\mathbf{y}}_{jfn}$. Note that if the source separation is needed as well, it can be immediately achieved by simply keeping the source images estimates.

2.3.2. *Model estimation*

Model parameters are estimated via a GEM algorithm [19] that is based on multiplicative update (MU) rules [21]. This is an iterative procedure, where the expectation step consists in computing the conditional expectations of empirical latent source image covariances $\widehat{\mathbf{C}}_{\mathbf{y}_{jfn}\mathbf{y}_{jfn}} \triangleq \mathbb{E}\left[\mathbf{y}_{jfn}\mathbf{y}_{jfn}^H \,|\bar{\mathbf{x}}'_n; \boldsymbol{\theta}\right]$, given the observations and the model, as

$$\widehat{\mathbf{C}}_{\mathbf{y}_{jfn}\mathbf{y}_{jfn}} = \hat{\mathbf{y}}_{jfn}\hat{\mathbf{y}}_{jfn}^H + \widehat{\boldsymbol{\Sigma}}_{\mathbf{y}_{jfn}\mathbf{y}_{jfn}}, \tag{10}$$

where $\hat{\mathbf{y}}_{jfn}$ and $\widehat{\boldsymbol{\Sigma}}_{\mathbf{y}_{jfn}\mathbf{y}_{jfn}}$ are computed as in (5) and (6).

Model parameters are then updated during maximization step as

$$\mathbf{R}_{jf} = \frac{1}{N}\sum_n \frac{1}{v_{jfn}}\widehat{\mathbf{C}}_{\mathbf{y}_{jfn}\mathbf{y}_{jfn}}, \tag{11}$$

$$\hat{p}_{jfn} = \frac{1}{I}\mathrm{tr}\left[\mathbf{R}_{jf}^{-1}\widehat{\mathbf{C}}_{\mathbf{y}_{jfn}\mathbf{y}_{jfn}}\right], \tag{12}$$

and

$$q_{jk} \leftarrow q_{jk}\left(\frac{\sum_{f,n} w_{fk}h_{nk}\hat{p}_{jfn}v_{jfn}^{-2}}{\sum_{f,n} w_{fk}h_{nk}v_{jfn}^{-1}}\right), \tag{13}$$

$$w_{fk} \leftarrow w_{fk}\left(\frac{\sum_{j,n} h_{nk}q_{jk}\hat{p}_{jfn}v_{jfn}^{-2}}{\sum_{j,n} h_{nk}q_{jk}v_{jfn}^{-1}}\right), \tag{14}$$

$$h_{nk} \leftarrow h_{nk}\left(\frac{\sum_{j,f} w_{fk}q_{jk}\hat{p}_{jfn}v_{jfn}^{-2}}{\sum_{j,f} w_{fk}q_{jk}v_{jfn}^{-1}}\right), \tag{15}$$

where equations (13) - (15) are MU rules minimizing the IS divergence [21] between tensors $\widehat{\mathbf{P}} = [\hat{p}_{jfn}]_{j,f,n}$ and $\mathbf{V} = [v_{jfn}]_{j,f,n}$ as in [14, 15].

2.4. Handling clipping constraint

In clipping, the original unknown signal is known to have its magnitude above clipping threshold outside the OS, and so should have the reconstructed signal frames. However, this constraint is not straightforward to handle within the above-described modeling, since the posterior distribution of source images (i.e., $\mathbf{y}_{jfn}$ given $\bar{\mathbf{x}}'_n$ and $\boldsymbol{\theta}$) is no longer Gaussian under this constraint. We have found in [9] that the following approach (referred to as *covariance projection* in [9]) is quite efficient. The following is done for each frame n. If after Wiener filtering (5) at least one windowed time domain sample in $\{\hat{x}'_{imn}\}_{i,m}$ does not satisfy the constraint, all the estimated samples not obeying the constraint are projected on the corresponding clipped signal and start to be considered as observed, i.e., their indices are added to the OS. Then, the Wiener filtering is recomputed over and over by applying the same strategy. This iterative process stops once all reconstructed samples obey the constraint. Once it is done, the GEM algorithm goes its usual way. Note that the above constraint handling process is re-started from the beginning at each GEM iteration, i.e., the OS is reset to its initial state at the beginning of each iteration. Here we apply exactly the same strategy, since its adaptation to the multichannel case does not present any difficulty.

2.5. Generalizing single-channel approaches and blind source separation

Note that the proposed multichannel framework including both the modeling and the GEM algorithm generalizes single-channel NMF/NTF-based frameworks [9, 18] we previously proposed. More precisely, for $I = 1$ and $J = 1$ the multichannel framework coincides exactly with the NMF-based single-channel declipping [9] up to some trivial component-wise scaling ($1 \times K$ matrix $\mathbf{Q}$) and some trivial frequency-wise scaling (1×1 matrices $\mathbf{R}_{1f}$). Moreover, for $I = 1$ it coincides exactly with the NTF-based single-channel joint declipping and source separation approach [18] up to a trivial source and frequency-wise scaling (1×1 matrices $\mathbf{R}_{jf}$). Finally, it is interesting to remark that when there is no clipping, i.e., $\overline{\Xi}'' = \emptyset$, the proposed multichannel framework reduces to a blind source separation (BSS) approach that is quite similar to, though not exactly coincides with, the BSS methods described in [22, 23].

3. EXPERIMENTS

3.1. Dataset

Since so far there is no available dataset for mutichannel declipping, we have created a small one ourselves using four convolutive stereo ($I = 2$) mixtures of three sources ($J^* = 3$)[4] from the `dev1` development dataset of the "Determined and over-determined speech and music mixtures" task of the 2010 Signal Separation Evaluation Campaign (SiSEC2010) [24]. To cover various cases we have chosen both speech and music signals, live and synthetic convolutive mixtures, different reverberation times (130 ms and 250 ms), and different microphone spacings (1 m and 5 cm). More specifically, we have chosen the following sequences (mixtures):

- Sequence 1: `dev1_female3_liverec_130ms_1m`
- Sequence 2: `dev1_male3_synthconv_130ms_5cm`
- Sequence 3: `dev1_nodrums_synthconv_250ms_1m`
- Sequence 4: `dev1_wdrums_liverec_250ms_5cm`

For each 10 sec length sequence sampled at 16 kHz we kept only the first 5 sec length half. Then, following [6], each multichannel signal was scaled to have maximum amplitude of 1 in time domain. Finally, each signal was clipped at clipping thresholds 0.7 and 0.9, which corresponds to a rather moderate clipping.[5]

3.2. Evaluation metric

To evaluate the performance we use the SNR_m metric [6], i.e., the signal to noise ratio (SNR) computed only on the clipped part. This metric was used to evaluate single-channel declipping [6, 9], but it extends very easily to the multichannel case as follows:

$$\mathrm{SNR}_m = 10\log_{10}\frac{\sum_{i=1}^{I}\|\mathbf{x}''_{i,\mathrm{orig}}(\bar{\Xi}''_i)\|^2}{\sum_{i=1}^{I}\|\mathbf{x}''_{i,\mathrm{orig}}(\bar{\Xi}''_i) - \mathbf{x}''_{i,\mathrm{est}}(\bar{\Xi}'')\|^2}, \tag{16}$$

[4] We denote the true number of sources in the mixture as J^*, since in this work we make a distinction between the true number of sources J^* and the number of sources J in the model.

[5] The choice of keeping only 5 second length signals and considering only moderate clipping is related to the fact that the computational load of the proposed approach is quite heavy so far. In case of moderate clipping a quite small portion of frames is really clipped, and for non-clipped frames our approach may be very efficiently optimized.

Metric & Method		Clipping level = 0.9					Clipping level = 0.7				
		Seq. 1	*Seq. 2*	*Seq. 3*	*Seq. 4*	*Average*	*Seq. 1*	*Seq. 2*	*Seq. 3*	*Seq. 4*	*Average*
$\mathrm{SNR_m}$ improvement (dB)	$J = 1$	1.30	3.84	2.89	0.58	**2.15**	3.34	-3.60	0.23	4.77	**1.18**
for the proposed	$J = 2$	3.12	9.20	2.70	-0.86	**3.54**	4.17	0.85	0.01	2.87	**1.98**
multichannel algorithm	$J = 3$	3.80	2.43	2.81	-0.18	**2.21**	7.13	-4.68	1.83	1.86	**1.53**
over the single-channel	$J = 4$	-0.89	-8.50	1.57	-0.28	**-2.02**	5.04	-3.77	-0.54	2.00	**0.68**
algorithm [9]	$J = 6$	-6.16	-19.84	8.52	1.84	**-3.90**	3.34	0.20	-0.78	1.45	**1.05**
$\mathrm{SNR_m}$ improvement (dB) for the single-channel algorithm [9] over the clipped signal		*17.83*	*34.42*	*12.66*	*7.67*	***18.15***	*16.09*	*17.71*	*17.06*	*9.99*	***15.21***

Table 1. SNR_m improvement (dB) for the proposed multichannel algorithm (for $J = 1, 2, 3, 4$ and 6) over the single-channel NMF-based algorithm [9] applied to each channel independently (top rows). SNR_m improvement (dB) for the single-channel algorithm [9] over the clipped signal (bottom row).

where $\mathbf{x}''_{i,\text{orig}}$ is the i-th channel of the original time domain signal, $\mathbf{x}''_{i,\text{est}}$ is the i-th channel of the estimated signal, $\bar{\Xi}''_i = \{1, \ldots, T\} \setminus \Xi''_i$ (with $\Xi''_i = \{t|(i,t) \in \Xi''\}$) is the set of time indices where the signal is lost due to clipping in the i-th channel, and $\|.\|$ denotes the ℓ_2 norm of a vector.

3.3. Parameters

For the proposed approach, the STFT is computed using a half-overlapping sine window of 1024 samples (64 ms) and the proposed GEM algorithm is run for 50 iterations. Following [9] the total number of components, K, is set to 20 for music signals and 28 for speech signals. For comparison, declipping on each channel is performed independently with the NMF based declipping algorithm in [9] with the number of components, STFT parameters and number of iterations being the same as the proposed approach. The proposed multichannel declipping algorithm is used with number of sources set to $J = 1, 2, 3, 4, 6$ in order to observe the change of performance with respect to the number of sources within the model. Before running the GEM algorithms all the model parameters are initialized with random values while assuring that the entries of $\mathbf{Q}$, $\mathbf{W}$ and $\mathbf{H}$ are all nonnegative and the covariance matrices $\mathbf{R}_{jf}$ are all positive definite and Hermitian.

3.4. Simulation results

The declipping performance of the proposed algorithm for different numbers of sources is presented in Table 1 along with the baseline result of declipping each channel independently. The results of the proposed algorithm is shown in terms of how much quality (measured in SNR_m) is increased with respect to the baseline result, hence positive values represent an improvement while negative values represent a decrease in quality.

The first thing to observe from the results in Table 1 is that at small J (1 or 2), the performance of the proposed approach is consistently better in average than the baseline. This is expected since the multichannel declipping algorithm exploits the correlation among the two channels and therefore provides a better approximation. As the number of sources increases however, the performance can be seen to be unstable, worse than the baseline more frequently. This is due to the fact that when performing multichannel declipping one must approximate the spatial covariance matrices, $\{\mathbf{R}_{jf}\}_{jf}$, and as the number of sources increases, the accurate estimation of correlation matrices gets increasingly harder and more sensitive to initialization. This is in line with the application of similar multichannel models and algorithms to convolutive BSS [12], where it is observed that the algorithms are very sensitive to initialization of parameters. More surprisingly this even affects the case $J = J^*$, so when the number of sources is correctly set in the algorithm, the performance is not always necessarily better than the performance with smaller J. However, the best average performance for both clipping levels (0.9 and 0.7) is achieved with $J = 2$, which proves the importance of considering a multi-source modeling for multichannel declipping. Finally, whatever the value of J, there is always at least one test sequence in our simulations for which the performance of the proposed multichannel algorithm is worth than that of its single-channel counterpart. In our opinion this may be related to the above mentioned sensitivity of the proposed multichannel algorithm to parameters initialisation. As such, we believe that initializing the model parameters with a more hand-crafted (not random) initialization may lead to a better declipping performance, as it is observed for convolutive BSS [12].

4. CONCLUSION

In this paper we have presented an audio declipping algorithm that can perform declipping on multichannel audio by exploiting the correlations between the channels in addition to the low rank NTF structure of the power spectrum of the audio signals. The proposed algorithm not only takes into account the correlations among the channels, but also the multi-source structure of the audio signals as well. It is shown that the proposed algorithm performs better than simply performing declipping on each channel when the number of sources in the estimation model is kept small. It is seen that when a larger number of sources is considered, the performance of the proposed method may drop due to large number of parameters to be estimated. Finally, while outperforming the baseline in average, the proposed approach does not always show a consistent improvement for all test sequences. We believe that this is due to its sensitivity to the parameters initialization.

Future work will include investigation of hand-crafted model parameters initialization strategies, investigation of the proposed approach in the context of joint declipping and convolutive BSS, as well as a more efficient implementation to enable faster estimation.

5. REFERENCES

[1] A. Adler, V. Emiya, M. Jafari, M. Elad, R. Gribonval, and M. D. Plumbley, "Audio inpainting," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 20, no. 3, pp. 922 – 932, 2012.

[2] A. Janssen, R. Veldhuis, and L. Vries, "Adaptive interpolation of discrete-time signals that can be modeled as autoregressive processes," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 34, no. 2, pp. 317–330, 1986.

[3] S. Abel and J.O. Smith III, "Restoring a clipped signal," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, 1991, p. 1745171748.

[4] S. J. Godsill and P. J. Rayner, "A Bayesian approach to the restoration of degraded audio signals," *IEEE Transactions on Speech and Audio Processing*, vol. 3, no. 4, pp. 267–278, 1995.

[5] S. Kitić, L. Jacques, N. Madhu, M.P. Hopwood, A. Spriet, and C. De Vleeschauwer, "Consistent iterative hard thresholding for signal declipping," in *ICASSP - The 38th International Conference on Acoustics, Speech, and Signal Processing*, Vancouver, Canada, May 2013.

[6] K. Siedenburg, M. Kowalski, and M. Dörfler, "Audio declipping with social sparsity," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, May 2014, pp. 1577–1581.

[7] M. J. Harvilla and R. M. Stern, "Efficient audio declipping using regularized least squares," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2015)*, Brisbane, Australia, Apr. 2015, pp. 221 – 225.

[8] S. Kitić, N. Bertin, and R. Gribonval, "Sparsity and cosparsity for audio declipping: a flexible non-convex approach," in *The 12th International Conference on Latent Variable Analysis and Signal Separation (LVA/ICA 2015)*, August, 2015.

[9] Ç. Bilen, A. Ozerov, and P. Pérez, "Audio declipping via nonnegative matrix factorization," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, October 2015.

[10] Xuemei Zhang and David H. Brainard, "Estimation of saturated pixel values in digital color imaging," *J. Opt. Soc. Am. A*, vol. 21, no. 12, pp. 2301–2310, 2004.

[11] H. Mansour, R. Saab, P. Nasiopoulos, and R. Ward, "Color image desaturation using sparse reconstruction," in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP 2010)*, Dallas, TX, USA, Mar. 2010.

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

[13] H. Sawada, S. Araki, and S. Makino, "Underdetermined convolutive blind source separation via frequency bin-wise clustering and permutation alignment," *IEEE Trans. Audio, Speech, Language Process.*, vol. 19, no. 3, pp. 516–527, 2011.

[14] C. Févotte and A. Ozerov, "Notes on nonnegative tensor factorization of the spectrogram for audio source separation : statistical insights and towards self-clustering of the spatial cues," in *7th International Symposium on Computer Music Modeling and Retrieval (CMMR 2010)*, 2010.

[15] A. Ozerov, C. Févotte, R. Blouet, and J.-L. Durrieu, "Multichannel nonnegative tensor factorization with structured constraints for user-guided audio source separation," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'11)*, Prague, May 2011, pp. 257–260.

[16] N. Q. K. Duong, E. Vincent, and R. Gribonval, "Underdetermined reverberant audio source separation using a full-rank spatial covariance model," *IEEE Trans. Audio, Speech, Language Process.*, vol. 18, no. 7, pp. 1830–1840, 2010.

[17] A. Liutkus, R. Badeau, and G. Richard, "Low bitrate informed source separation of realistic mixtures," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Vancouver, Canada, May 2013.

[18] Ç. Bilen, A. Ozerov, and P. Pérez, "Joint audio inpainting and source separation," in *The 12th International Conference on Latent Variable Analysis and Signal Separation (LVA/ICA 2015)*, August 2015.

[19] A.P. Dempster, N.M. Laird, and D.B. Rubin., "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 39, pp. 1–38, 1977.

[20] S. M. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory*, Prentice Hall, Englewood Cliffs, NJ, 1993.

[21] C. Févotte, N. Bertin, and J.L. Durrieu, "Nonnegative matrix factorization with the Itakura-Saito divergence. With application to music analysis," *Neural Computation*, vol. 21, no. 3, pp. 793–830, Mar. 2009.

[22] S. Arberet, A. Ozerov, N.Q.K. Duong, E. Vincent, R. Gribonval, F. Bimbot, and P. Vandergheynst, "Nonnegative matrix factorization and spatial covariance model for underdetermined reverberant audio source separation," in *10th Int. Conf. on Information Sciences, Signal Proc. and their applications (ISSPA1710)*, 2010, pp. 1–4.

[23] A. Ozerov, E. Vincent, and F. Bimbot, "A general flexible framework for the handling of prior information in audio source separation," *IEEE Trans. Audio, Speech, Language Process.*, vol. 20, no. 4, pp. 1118–1133, 2012.

[24] S. Araki, A. Ozerov, B.V. Gowreesunker, H. Sawada, F.J. Theis, G. Nolte, D. Lutter, and N.Q.K. Duong, "The 2010 signal separation evaluation campaign (SiSEC2010): - Audio source separation," in *Proc. Int. Conf. on Latent Variable Analysis and Signal Separation*, 2010, pp. 114–122.