Audio declipping via nonnegative matrix factorization

Cagdas Bilen, Alexey Ozerov, Patrick Pérez

AUDIO DECLIPPING VIA NONNEGATIVE MATRIX FACTORIZATION

Çağdaş Bilen, Alexey Ozerov* and Patrick Pérez*

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

ABSTRACT

Audio inpainting and audio declipping are important problems in audio signal processing, which are encountered in various practical applications. A number of approaches has been proposed in the literature to address these problems, most successful of which are based on sparsity of the audio signals in certain dictionary representations. Non-negative matrix factorization (NMF) is another powerful tool that has been successfully used in applications such as audio source separation. In this paper we propose a new algorithm that makes use of a low rank NMF model to perform audio inpainting and declipping. In addition to utilizing for the first time the NMF model to perform audio inpainting in presence of arbitrary losses in time domain, the proposed approach also introduces a novel way to enforce additional constraints on the signal magnitude in order to improve the performance in declipping applications. The proposed approach is shown to have a comparable performance with the state of the art dictionary based methods while providing a number of advantages.

Index Terms— Audio inpainting, audio declipping, nonnegative matrix factorization, Itakura-Saito divergence, generalized expectation-maximization

1. INTRODUCTION

Different problems that consist in recovering some missing parts of an audio signal (e.g., packet loss concealment or bandwidth expansion) were recently regrouped and baptized *audio inpainting* [1], by analogy with image inpainting [2]. In this paper we consider a particular audio inpainting problem when some small chunks or samples of an audio signal are missing in the time domain. This happens in practice at least in two situations. First, when the audio signal is clipped above a certain amplitude, and in this case one speaks about *audio declipping* [3]. Second, when some portions of the original audio signal are corrupted by impulse noise or "clicks", and this problem is referred to as *click removal* [4]. The only difference between these two problems is that in the former one there is a little bit more knowledge about missing audio samples than in the latter one. Indeed, in the case of audio declipping the missing samples must be either above or below the corresponding clipping threshold, which we here refer to as *clipping constraint* [5]. In this paper, we focus on the audio declipping problem and evaluate the proposed approach on this task. However, the approach we propose works as well without clipping constraints and can thus be applied also to click removal or to other audio inpainting problems in which arbitrary parts of time domain signals are missing.

* The first and second authors have contributed equally for this work.

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

Audio declipping and click removal were addressed in the past by autoregressive (AR) modeling [6], signal matching with bandwidth constraints [3], and Bayesian estimation [4]. Adler *et al.* [1] have formulated audio declipping and click removal as inverse problems and addressed them using a sparse decomposition approach based on the orthogonal matching pursuit (OMP) algorithm. They obtained results that are on a par with state-of-the-art audio declipping [6]. Since the publishing of [1], the problem of audio declipping has regained interest and several new improved methods were proposed. Kitić *et al.* proposed a sparse decomposition approach based on the iterative hard-thresholding (HT) [7], which was extended then to cosparse decomposition [8]. Siedenburg *et al.* [5] proposed an approach based on social sparsity [9], which is a new variant of structured sparsity. All these recent methods are based either on sparsity alone [1, 7, 8] or on a model representing some local structure of an audio signal [5].

The goal of our work is to propose and investigate a model representing the global rather than the local structure of the audio signal. To this end nonnegative matrix factorization (NMF) [10], which has recently found great success in audio source separation [11, 12, 13], audio compression [14, 15], music transcription [16, 17] and some forms of audio inpainting [18, 19, 20, 21], is an appealing tool. Indeed, it is an object-based decomposition [10] that allows modeling similarities between audio patterns within an audio signal [12]. Although the NMF was already applied to audio inpainting [19, 20, 21], it has not yet been used to recover time domain losses with arbitrary loss patterns and its application to audio declipping remains challenging. Indeed, the model being defined on some time-frequency representation, usually the short-time Fourier transform (STFT) domain as we consider hereafter, the state-of-the-art approaches [19, 20, 21] assume that the data losses occur in the same domain. For audio declipping however, the missing data are small chunks in the time domain and it is in general impossible to convert them into the STFT domain without a considerable loss of the information.

In this work we propose a practical audio declipping approach that relies on the NMF model in the STFT domain and exploits at the same time all available observed information in the time domain. We make it possible by resorting to the Itakura Saito (IS) NMF [12]. Thanks to its probabilistic formulation [12] as a Gaussian distribution of complex-valued STFT coefficients, it becomes possible to switch within this framework between the time and the STFT domains. We estimate the NMF model from the partial observations in the maximum likelihood (ML) sense, thanks to a new generalized expectation-maximization (GEM) algorithm [22] based on the multiplicative update (MU) rules [12]. Once the NMF model is estimated, the missing audio samples may be predicted using the Wiener filter [23]. Yet, handling the clipping constraint remains difficult,since it is specified in the time domain, while the modeling is

defined in the STFT domain. We propose a practical solution to circumvent this problem.

The rest of this paper is organized as follows. The formal definition of the audio inpainting/declipping problem and the assumed signal model are introduced in Section 2. The proposed GEM algorithm is presented in Section 3 with a discussion on how to handle the clipping constraint. Results and comparison with the state-of-the-art methods [1, 7, 5] are reported in Section 4, and some conclusions are drawn in Section 5.

2. PROBLEM FORMULATION

2.1. Definitions and basic setup

We consider the single channeltime domain signal

$$x''_t = s''_t + a''_t, \quad t = 1 \cdots T, \tag{1}$$

where $t = 1, \ldots, T$ is the discrete time index, and x''_t, s''_t, and a''_t denote respectively measured signal, original source and quantization and/or measurement noise samples.[1] Moreover, it is assumed that the audio signal is only observed on a subset of time indices $\Xi'' \subset \{1, \ldots, T\}$ called *observation support (OS)*. For clipped signals this support indicates the indices of time instants where signal magnitude is smaller than the clipping threshold. In the rest of this paper we will assume that there is no noise in the measured signal, i.e. $a''_t = 0$, for simplicity. The proposed algorithm can be easily extended to work with the noisy case. Furthermore it is also assumed that the support of the signal is known and if the signal is clipped, the clipping threshold is known.

In the windowed time domain, the time domain signals are taken in overlapping (often non-rectangular) frames of length M. Measured and original audio signals become $\{x'_{mn}\}$ and $\{s'_{mn}\}$ 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 $\Xi' \subset \{1, \ldots, M\} \times \{1, \ldots, N\}$ whose restriction to n-th frame reads $\Xi'_n = \{m | (m, n) \in \Xi'\}$. The Short-Time Fourier Transfrom (STFT) coefficients of the time domain signals x''_t and s''_t 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 $\mathbf{x}_n = \mathbf{U}\mathbf{x}'_n$ and $\mathbf{s}_n = \mathbf{U}\mathbf{s}'_n$ where each vector represents the samples within a frame, for example $\mathbf{x}'_n = [x'_{mn}]_{m=1\cdots M}$ and $\mathbf{x}_n = [x_{fn}]_{f=1\cdots F}$, with index f representing the frequency index of the Fourier transform coefficients.

In this paper, we consider the specific problem of clipping, which implies additional constraints. However the proposed algorithm can easily be applied the generic audio inpainting problem by omitting the steps specifically dedicated to the clipping constraints. In the rest of this paper, we will denote the observed clipped signal in windowed time domain as $\mathbf{x}'_c$ and its restriction to un-clipped instants as $\bar{\mathbf{x}}'$, where $\bar{\mathbf{x}}'_n = [x'_{mn}]_{m \in \Xi'_n}$.

2.2. Model

The original audio signal is modelled in the STFT domain with a normal distribution ($s_{fn} \sim \mathcal{N}_c(0, v_{fn})$) where the variance matrix $\mathbf{V} = [v_{fn}]$ has the following low-rank NMF structure [24, 15]

$$v_{fn} = \sum_{k=1}^{K} w_{fk} h_{nk}, \tag{2}$$

where K is small and all the variables are non-negative reals. This model amounts to $\mathbf{V} = \mathbf{W}\mathbf{H}^\top$ and is parametrized by $\boldsymbol{\theta} = \{\mathbf{W}, \mathbf{H}\}$, with $\mathbf{W} = [w_{fk}]_{f,k}$ and $\mathbf{H} = [h_{nk}]_{n,k}$ being, respectively, $F \times K$ and $N \times K$ non-negative matrices.

[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.

Algorithm 1 GEM algorithm for NMF model estimation

procedure INPAINTING-NMF($\mathbf{x}'_c, \Xi', \bar{\mathbf{x}}'$)
  Initialize non-negative $\mathbf{W}, \mathbf{H}$ randomly
  repeat
    Estimate $\hat{\mathbf{s}}$ (orig. signal), given $\mathbf{W}, \mathbf{H}, \bar{\mathbf{x}}', \Xi'$ ▷ E-step, see section 3.1
    Estimate $\tilde{\mathbf{s}}$ (signal obeying clipping constraint) and $\tilde{\mathbf{P}}$ (posterior power spectra), given $\hat{\mathbf{s}}$, $\mathbf{W}, \mathbf{H}, \bar{\mathbf{x}}', \Xi'$ and $\mathbf{x}'_\mathbf{c}$ ▷ Applying clipping constraint, see section 3.2
    Update $\mathbf{W}, \mathbf{H}$ given $\tilde{\mathbf{P}}$ ▷ M-step, see section 3.3
  until convergence criteria met
end procedure

3. MODEL ESTIMATION AND SIGNAL RECONSTRUCTION

We propose to estimate model parameters using a generalized expectation-maximization (GEM) algorithm [22] and to estimate the signal using Wiener filtering [23]. The proposed algorithm is outlined in Algorithm 1, and its steps described below.

3.1. Estimation of the signal

All the underlying distributions are assumed to be Gaussian and all the relations between the source signal and the observations are linear, except the clipping constraint that will be addressed specifically in Section 3.2. Thus, without taking into account the clipping constraint, the source can be estimated in the minimum mean square error (MMSE) sense via Wiener filtering [23] given the covariance matrix $\mathbf{V}$ defined in (2) by the model parameters $\mathbf{W}, \mathbf{H}$.

We can write the posterior distribution of each signal frame $\mathbf{s}_n$ given the corresponding observed frame $\bar{\mathbf{x}}'_n$ and NMF model $\boldsymbol{\theta}$ as $\mathbf{s}_n | \bar{\mathbf{x}}'_n; \boldsymbol{\theta} \sim \mathcal{N}_c(\hat{\mathbf{s}}_n, \widehat{\boldsymbol{\Sigma}}_{\mathbf{s}_n\mathbf{s}_n})$ with $\hat{\mathbf{s}}_n$ and $\widehat{\boldsymbol{\Sigma}}_{\mathbf{s}_n\mathbf{s}_n}$ being, respectively, posterior mean and posterior covariance matrix, each of which can be computed by Wiener filtering as[2]

$$\hat{\mathbf{s}}_n = \boldsymbol{\Sigma}^H_{\bar{\mathbf{x}}'_n\mathbf{s}_n} \boldsymbol{\Sigma}^{-1}_{\bar{\mathbf{x}}'_n\bar{\mathbf{x}}'_n} \bar{\mathbf{x}}'_n, \tag{3}$$

$$\widehat{\boldsymbol{\Sigma}}_{\mathbf{s}_n\mathbf{s}_n} = \boldsymbol{\Sigma}_{\mathbf{s}_n\mathbf{s}_n} - \boldsymbol{\Sigma}^H_{\bar{\mathbf{x}}'_n\mathbf{s}_n} \boldsymbol{\Sigma}^{-1}_{\bar{\mathbf{x}}'_n\bar{\mathbf{x}}'_n} \boldsymbol{\Sigma}_{\bar{\mathbf{x}}'_n\mathbf{s}_n}, \tag{4}$$

given the definitions

$$\boldsymbol{\Sigma}_{\mathbf{s}_n\mathbf{s}_n} \triangleq \mathrm{diag}\left([v_{fn}]_f\right), \boldsymbol{\Sigma}_{\bar{\mathbf{x}}'_n\mathbf{s}_n} \triangleq \mathbf{U}(\Xi'_n)^H \mathrm{diag}\left([v_{fn}]_f\right), \tag{5}$$

$$\boldsymbol{\Sigma}_{\bar{\mathbf{x}}'_n\bar{\mathbf{x}}'_n} \triangleq \mathbf{U}(\Xi'_n)^H \mathrm{diag}\left([v_{fn}]_f\right) \mathbf{U}(\Xi'_n), \tag{6}$$

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

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

3.2. Handling clipping constraint

In order to update the model parameters as will be described in Section 3.3, one needs to estimate the posterior power spectra of the signal defined as

$$\tilde{p}_{fn} = \mathbb{E}\left[\,|s_{fn}|^2 \,\middle|\, \bar{\mathbf{x}}'_n; \boldsymbol{\theta}\right]. \tag{7}$$

For an audio inpainting problem without any further constraints, the posterior signal estimate, $\hat{\mathbf{s}}_n$, and the posterior covariance matrix, $\widehat{\boldsymbol{\Sigma}}_{\mathbf{s}_n\mathbf{s}_n}$, computed as in Section 3.1 would be sufficient to estimate $\tilde{p}_{fn}$, since the posterior distribution of the signal is Gaussian. However 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 $\hat{\mathbf{s}}'_n = \mathbf{U}^H\hat{\mathbf{s}}_n$:

$$\hat{s}'_{mn} \times \operatorname{sign}\left(x'_{mn}\right) \geq \left|x'_{mn}\right|, \; \forall n, \; \forall m \notin \Xi'_n. \tag{8}$$

This constraint is difficult to enforce directly into the model since the posterior distribution of the signal under it is no longer Gaussian, which significantly complicates the computation of the posterior power spectra. In the presence of such constraints on the magnitude of the signal, we consider various ways to approach the problem:

1. **Unconstrained:** The simplest way to perform the estimation is to ignore completely the constraints, treating the problem as a more generic audio inpainting in time domain. Hence during the iterations, the "constrained" signal is taken simply as the estimated signal, i.e. $\tilde{\mathbf{s}}_n = \hat{\mathbf{s}}_n, n = 1, \ldots, N$, as is the posterior covariance matrix, $\widetilde{\boldsymbol{\Sigma}}_{\mathbf{s}_n\mathbf{s}_n} = \widehat{\boldsymbol{\Sigma}}_{\mathbf{s}_n\mathbf{s}_n}, n = 1, \ldots, N$.
2. **Ignored projection:** Another simple way to proceed is to ignore the constraint during the iterative estimation process and to enforce it at the end as a post-processing of the estimated signal. In this case, the signal is treated the same as the unconstrained case during the iterations.
3. **Signal projection:** A more advanced approach is to update the estimated signal at each iteration so that the magnitude obeys the clipping constraints. Let's suppose (8) is not satisfied at the indices in set $\hat{\Xi}'_n$. We can set $\tilde{\mathbf{s}}'_n = \hat{\mathbf{s}}'_n$ and then force $\tilde{\mathbf{s}}'_n(\hat{\Xi}'_n) = \mathbf{x}'_{c,n}(\hat{\Xi}'_n)$. However this approach does not update the posterior covariance matrix, i.e. $\widetilde{\boldsymbol{\Sigma}}_{\mathbf{s}_n\mathbf{s}_n} = \widehat{\boldsymbol{\Sigma}}_{\mathbf{s}_n\mathbf{s}_n}, n = 1, \ldots, N$.
4. **Covariance projection:** In order to update as well the posterior covariance matrix, we can re-compute the posterior mean and the posterior covariance by (3) and (4) respectively, using $\Xi'_n \cup \hat{\Xi}'_n$ instead of Ξ'_n and $\mathbf{x}'_{c,n}(\Xi'_n \cup \hat{\Xi}'_n)$ instead of $\bar{\mathbf{x}}'_n$ in equations (3)-(6). If the resulting estimation of the sources violate (8) on additional indices, $\hat{\Xi}'_n$ is extended to include these indices and the computation is repeated. As a result, the final signal estimate, $\tilde{\mathbf{s}}$, which satisfies (8) and the corresponding posterior covariance matrix, $\widetilde{\boldsymbol{\Sigma}}_{\mathbf{s}_n\mathbf{s}_n}$, are obtained. Note that in addition to updating the posterior covariance matrix, this approach also updates the entire estimated signal and not just the signal at the indices of violated constraints.

After using one of the above approaches, the posterior power spectra, $\tilde{\mathbf{P}} = [\tilde{p}_{fn}]$ can be computed as

$$\tilde{p}_{fn} = \mathbb{E}\left[\,|s_{fn}|^2 \,\middle|\, \bar{\mathbf{x}}'_n; \boldsymbol{\theta}\right] \cong |\tilde{s}_{fn}|^2 + \widetilde{\boldsymbol{\Sigma}}_{\mathbf{s}_n\mathbf{s}_n}(f,f). \tag{9}$$

3.3. Updating the model

NMF model parameters can be re-estimated using the multiplicative update (MU) rules minimizing the IS divergence [12] between the matrix of estimated signal power spectra $\tilde{\mathbf{P}} = [\tilde{p}_{fn}]_{f,n}$ and the NMF model approximation $\mathbf{V} = \mathbf{W}\mathbf{H}^\top$:

$$D_{IS}(\tilde{\mathbf{P}}\|\mathbf{V}) = \sum_{f,n} d_{IS}(\tilde{p}_{fn}\|v_{fn}), \tag{10}$$

where $d_{IS}(x\|y) = x/y - \log(x/y) - 1$ is the IS divergence, and $\tilde{p}_{fn}$ and v_{fn} are specified respectively by (9) and (2). Hence the model parameters can be updated as

$$w_{fk} \leftarrow w_{fk}\left(\frac{\sum_n h_{nk}\tilde{p}_{fn}v_{fn}^{-2}}{\sum_n h_{nk}v_{fn}^{-1}}\right), \tag{11}$$

$$h_{nk} \leftarrow h_{nk}\left(\frac{\sum_f w_{fk}\tilde{p}_{fn}v_{fn}^{-2}}{\sum_f w_{fk}v_{fn}^{-1}}\right). \tag{12}$$

This update can also be repeated a number of times to improve the model prediction.

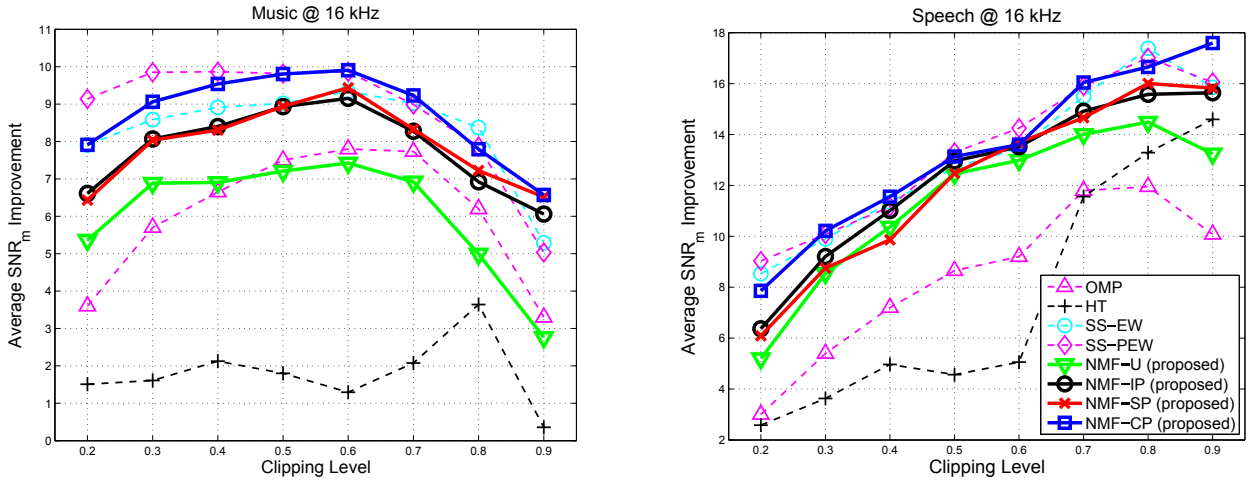
4. RESULTS

In order to measure the performance of the proposed algorithm, 10 music and 10 speech signals at sampling frequency of 16 kHz are artificially clipped with eight different clipping thresholds. The improvement of the signal to noise ratio (computed only on the clipped regions) with respect to the clipped signal, SNR_m, is computed as:

$$\text{SNR}_m\left(\mathbf{x}''_{\text{orig}}, \mathbf{x}''_{\text{est}}, \bar{\Xi}''\right) = 10\log_{10}\frac{\|\mathbf{x}''_{\text{orig}}(\bar{\Xi}'')\|^2}{\|\mathbf{x}''_{\text{orig}}(\bar{\Xi}'') - \mathbf{x}''_{\text{est}}(\bar{\Xi}'')\|^2}, \tag{13}$$

where $\mathbf{x}''_{\text{orig}}$ is the original time domain signal, $\mathbf{x}''_{\text{est}}$ is the estimated signal, $\bar{\Xi}'' = \{1 \cdots T\} \setminus \Xi''$ is the set of time indices where the signal is lost due to clipping and $\|.\|$ denotes the ℓ_2 norm of a vector. The methodology and the dataset used for the experiments are identical to the ones reported in [5]. Each audio signal of length 4 seconds sampled at 16 kHz is scaled to have maximum magnitude of 1 in time domain, and then clipped at different clipping thresholds (from 0.2 to 0.9). The proposed algorithm is tested with four different methods to handle clipping constraint as described in Section 3.2, namely *Unconstrained (NMF-U)*, *Ignored Projection (NMF-IP)*, *Signal Projection (NMF-SP)* and *Covariance Projection (NMF-CP)*. The music signals are declipped with 20 NMF components ($K = 20$) while 28 components are used for speech signals ($K = 28$). 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. The performance of the proposed algorithm is compared to four state of the art methods: iterative hard-thresholding (HT) [7], orthogonal matching pursuit (OMP) [1], social sparsity with empirical Wiener operator (SS-EW) and social sparsity with posterior empirical Wiener operator (SS-PEW) [5].

The average performance of all the algorithms for declipping of music and speech signals is represented on Figure 1. It can be seen from the overall results that the proposed algorithm with the covariance projection (NMF-CP) has almost identical performance with the social sparsity based methods (SS-EW and SS-PEW) proposed in [5] while outperforming others. It can be also seen in the results that the model based algorithms (social sparsity and the NMF



(a) Average SNR$_m$ improvement computed over 10 music signals. (b) Average SNR$_m$ improvement computed over 10 speech signals.

Figure 1: The average performance of all the algorithms as a function of the clipping threshold (lower threshod corresponds to more severe clipping).

model) significantly outperform the methods relying on just sparsity (OMP and HT).

Even though the performance is not better than the social sparsity approaches at first glance, the proposed algorithm has a number of advantages over the other methods:

- NMF model can be easily extended to other, more structured NMF-like models such as source-excitation model or harmonic NMF [25]. As shown in [25] in case of source separation, having a specific model with structure that is well adapted to the considered class of signal (e.g., speech, music, etc.) may improve the overall performance.
- It is known that the NMF model (or more general non-negative tensor factorization) can help greatly in other audio signal recovery problems such as source separation. Hence the presented algorithm also provides a flexible framework to perform other tasks such as joint source separation and declipping, which is investigated in [26].
- It is shown in the results that the performance of our method depends significantly one the way the clipping constraint is handled. Therefore an alternative, more accurate computation of the posterior power spectrum might also improve the results further, whereas in dictionary based methods there is no approximation for the clipping constraints, hence performance improvement in this regard is not possible.

Regarding the effect of clipping constraints, the first thing to notice is that the performance of NMF-U with respect to NMF-IP (and NMF-SP) shows that simple constraints on the signal magnitude can noticably improve the performance especially for music signals, hence they should not be ignored when possible. NMF-IP and NMF-SP are shown to have almost identical performance, even though the latter applies the constraints on the posterior mean of the signal at every iteration and the former simply applies a post processing to the final result. This observation combined with the superior performance of NMF-CP compared to the other methods demonstrates the importance of updating the posterior power spectrum more accurately for the success of the NMF-based methods.

It should be noted that dealing with constraints in both time domain and STFT domain comes at a computational cost in the Wiener filtering stage of the proposed algorithm. Luckily, this step is independent for each frame of the signal and hence can be easily parallelized, e.g., using graphical processing units (GPUs), to get significant speed-up. On the other hand, estimating the signal independently within each window comes with the disadvantage that the estimation is not possible when there are no observed samples within a window. In practice, however, the loss of an entire window due to clipping is not probable for natural audio signals when the window size is chosen properly and the clipping threshold is not extremely low.

5. CONCLUSION

In this paper, we introduce the first NMF-based technique to perform audio inpainting of signals with arbitrary loss patterns in time domain. Furthermore, a novel approach to deal with clipping constraints on time domain signal magnitude has been introduced within the proposed GEM algorithm. We showed that our NMF-based approach can be used to achieve state of the art audio declipping performance while giving the flexibility to embrace other forms of constraints on the signal and/or to perform additional tasks such as audio source separation.

Improved and more accurate methods to enforce the clipping constraints and other constraints in STFT domain are possible directions for future research. Investigating the performance of the proposed approach in the presence of noise as well as comparing it to other methods such as cosparse hard-thresholding [8] are also considered as future work. Lastly, another interesting direction for research is the extension of the proposed algorithm to handle multichannel audio.

6. ACKNOWLEDGEMENT

The authors would like to thank Kai Siedenburg and Matthieu Kowalski for kindly sharing numerical results from [5].

7. 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] M. Bertalmio, G. Sapiro, V. Caselles, and C. Ballester, "Image inpainting," in *SIGGRAPH'00*, 2000, pp. 417–424.

[3] S. Abel and J. Smith III, "Restoring a clipped signal," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, 1991, p. 17451748.

[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] 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.

[6] 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.

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

[8] S. Kitić, N. Bertin, and R. Gribonval, "Audio declipping by cosparse hard thresholding," in *iTwist - 2nd international - Traveling Workshop on Interactions between Sparse models and Technology*, Namur, Belgium, Aug. 2014.

[9] M. Kowalski, K. Siedenburg, and M. Dörfler, "Social sparsity! Neighborhood systems enrich structured shrinkage operators," *IEEE Transactions on Signal Processing*, vol. 61, no. 10, pp. 2498–2511, May 2013.

[10] D. Lee and H. Seung, "Learning the parts of objects with nonnegative matrix factorization," *Nature*, vol. 401, pp. 788–791, 1999.

[11] T. Virtanen, "Monaural Sound Source Separation by Nonnegative Matrix Factorization With Temporal Continuity and Sparseness Criteria," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 3, pp. 1066–1074, 2007.

[12] C. Févotte, N. Bertin, and J. 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.

[13] 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.

[14] J. Nikunen and T. Virtanen, "Object-based audio coding using non-negative matrix factorization for the spectrogram representation," in *128th Audio Engineering Society Convention (AES 2010)*, London, UK, May 2010.

[15] A. Ozerov, A. Liutkus, R. Badeau, and G. Richard, "Coding-based informed source separation: Nonnegative tensor factorization approach," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 8, pp. 1699–1712, Aug. 2013.

[16] P. Smaragdis and J. C. Brown, "Non-negative matrix factorization for polyphonic music transcription," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, Oct. 2003.

[17] E. Vincent, N. Bertin, and R. Badeau, "Adaptive harmonic spectral decomposition for multiple pitch estimation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 3, pp. 528–537, Mar. 2010.

[18] T. Virtanen, A. Mesaros, and M. Ryynnen, "Combining pitch-based inference and non-negative spectrogram factorization in separating vocals from polyphonic music," in *ISCA Tutorial and Research Workshop on Statistical And Perceptual Audition. SAPA 2008*, 2008.

[19] J. L. Roux, H. Kameoka, N. Ono, A. de Cheveigné, and S. Sagayama, "Computational auditory induction as a missing-data model-fitting problem with Bregman divergence," *Speech Communication*, vol. 53, no. 5, pp. 658–676, May-June 2011.

[20] P. Smaragdis, R. Bhiksha, and S. Madhusudana, "Missing data imputation for time-frequency representations of audio signals," *Journal of signal processing systems*, vol. 65, no. 3, pp. 361–370, 2011.

[21] U. Simsekli, A. T. Cemgil, and Y. K. Yilmaz, "Score guided audio restoration via generalised coupled tensor factorisation," in *International Conference on Acoustics Speech and Signal Processing (ICASSP'12)*, 2012, pp. 5369 – 5372.

[22] A. Dempster, N. Laird, and D. 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.

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

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

[25] 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.

[26] Ç. Bilen, A. Ozerov, and P. Pérez, "Joint audio inpainting and source separation," in *International Conference on Latent Variable Analysis and Signal Separation*, August 2015.