Joint Audio Inpainting and Source Separation

Cagdas Bilen, Alexey Ozerov, Patrick Pérez

Joint Audio Inpainting and Source Separation

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. Despite being two important problems in audio signal processing that are interconnected in practice, audio inpainting and audio source separation have not been considered jointly. It is not uncommon in practice to have the mixtures to be separated which also suffer from artifacts due to clipping or other losses. In present work, we consider this problem of source separation using partially observed mixtures. We introduce a flexible framework based on non-negative tensor factorisation (NTF) to attack this new task, and we apply it to source separation with clipped mixtures. It allows us to perform declipping and source separation either in turn or jointly. We investigate experimentally these two regimes and report large performance gains compared to source separation with clipping artefacts being ignored, which is the common approach in practice.

1 Introduction

Audio inpainting and audio source separation are two important problems in audio signal processing. The former is defined as the one of reconstructing the missing parts in an audio signal [1]. It's been coined "audio inpainting" to draw an analogy with visual inpainting, a widely studied problem where the goal is to reconstruct regions in images, for restoration or editing purposes [2]. We consider here the problem of audio inpainting in which some temporal audio samples are lost (as opposed to earlier works where losses are in time frequency domain), such as with saturation of amplitude (clipping) or interfering high amplitude impulsive noise (clicking), and need to be recovered (called declipping and declicking for these two specific cases respectively).

The problem of audio source separation is the one of separating an audio signal into meaningful, distinctive sources which add up to a known mixture, such as separating a music signal into signals from different instruments. Even though audio source separation and audio inpainting have been studied extensively, these two problems have not yet been considered jointly. There are common situations,

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

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

however, where one task should benefit from the other and vice versa: many audio signals to be de-clipped are in fact composed of multiple sources and, conversely, the audio mixtures in various source separation tasks might be clipped due to the nature of recording equipment. Hence considering these two tasks simultaneously could help improve the performance of both.

In this paper we propose a first approach toward this goal. To this end, we extend our recent work on audio inpainting with application to declipping [3]. This approach, based on non-negative matrix factorization (NMF) performs as well or better than the state of the art group sparsity based methods such as [10]. It builds on the recent successes of NMF [7] and non-negative tensor factorization (NTF) in audio inpainting [9, 11, 3] [1]. Since, NMF/NTF framework is also very powerful in source separation [13, 5, 8], it lends itself to addressing the joint problem of audio inpainting and audio source separation.

Extending [3], we estimate individual sources using a low rank NTF model, with the help of some temporal source activity information as in [8]. The proposed algorithm not only can perform audio inpainting and source separation sequentially (i.e., first inpaint the mixture, then separate the sources), but also can perform these two tasks jointly (i.e., simultaneously inpaint the mixture and separate the sources). It is shown that joint inpainting and separation benefits both tasks greatly, especially when the loss due to clipping is significant. The performance of both the sequential and the joint approaches are shown to be much better than the performance of source separation when the degradation due to clipping is ignored as it is usually the case in practice. Section 2 is devoted to problem formulation and modeling description. The main algorithm is outlined in Section 3. The experiments are presented in Section 4, and some conclusions are drawn in Section 5.

2 Signal Model and Problem Formulation

Let us consider the following single-channel[2] mixing equation in time domain:

$$x''_t = \sum_{j=1}^{J} s''_{jt} + a''_t, \quad t \in [\![1, T]\!], j \in [\![1, J]\!] \tag{1}$$

where t is the discrete time index, j is the source index, and x''_t, s''_{jt}, and a''_t denote respectively mixture, source and quantization noise samples.[3] It is assumed that

[1] As opposed to [3], earlier works on audio inpainting with NMF/NTF models [9, 11] cannot optimally address arbitrary losses in time domain, since the missing data are formulated in time frequency domain.

[2] This work would be readily extended to the multi-channel case. For sake of simplicity, we only consider the single-channel case here.

[3] Throughout, letters with two primes, e.g., x'', denote time domain signals, letters with one prime, e.g., x' denote framed and windowed-time domain signals and letters with no primes, e.g., x, denote complex-valued short-time Fourier transform (STFT) coefficients.

the mixture is only observed on a subset of time indices $\Xi'' \subset [\![1, T]\!]$ called *mixture observation support* (MOS). For clipped signals this support indicates the indices with signal magnitude smaller than the clipping threshold. For the rest of this paper, we assume for sake of simplicity that there is no mixture quantization ($a''_t = 0$) .

The sources are unknown. We assume, however, that it is known which sources are *active* at which time periods. For a multi-instrument music for instance, this information corresponds to knowing which instruments are playing at any instant. Furthermore it is also assumed that if the mixture is clipped, the clipping threshold is known.

In order to compute the STFT coefficents, the mixture and the sources are first converted to windowed-time domain with a window length M and a total of N windows with resulting coefficients denoted by s'_{jmn} and x'_{mn} representing the original sources and the mixture in windowed-time domain respectively for $m = [\![1, M]\!], n = [\![1, N]\!], j = [\![1, J]\!]$. We also introduce the set $\Xi' \subset [\![1, N]\!] \times [\![1, M]\!]$ that is the MOS within the framed representation corresponding to Ξ'' in time domain, and its frame-level restriction $\Xi'_n = \{m|(m, n) \in \Xi'\}$. We will denote the observed clipped mixture in windowed-time domain as $\mathbf{x}'_c = \{\mathbf{x}'_{c,n}\}_{n=1}^N$ and its restriction to un-clipped instants as $\bar{\mathbf{x}}' = \{\bar{\mathbf{x}}'_n\}_{n=1}^N$, where $\bar{\mathbf{x}}'_n = [x'_{mn}]_{m\in\Xi'_n}$. The STFT coefficients of the sources, s_{jfn}, and the mixture, x_{fn}, are computed via applying a unitary fourier transfrom, $\mathbf{U} \in \mathbb{C}^{F\times M} (F = M)$, to each window of the windowed-time domain counterparts. For example, $[x_{1n}, \cdots, x_{Fn}]^T = \mathbf{U}[x'_{1n}, \cdots, x'_{Mn}]^T$[4].

The sources are modelled in the STFT domain with a normal distribution ($s_{jfn} \sim \mathcal{N}_c(0, v_{jfn})$) where the variance tensor $\mathbf{V} = [v_{jfn}]_{j,f,n}$ has the low-rank NTF structure (with a small K) [8] such that $v_{jfn} = \sum_{k=1}^K q_{jk} w_{fk} h_{nk}$ with $q_{jk}, w_{fk}, h_{nk} \in \mathbb{R}_+$. This model is parametrized by $\boldsymbol{\theta} = \{\mathbf{Q}, \mathbf{W}, \mathbf{H}\}$, with $\mathbf{Q} = [q_{jk}]_{j,k} \in \mathbb{R}_+^{J\times K}$, $\mathbf{W} = [w_{fk}]_{f,k} \in \mathbb{R}_+^{F\times K}$ and $\mathbf{H} = [h_{nk}]_{n,k} \in \mathbb{R}_+^{N\times K}$.

The assumed information on which sources are active at which time periods is captured by constraining certain entries of $\mathbf{Q}$ and $\mathbf{H}$ to be zero as in [8]. Each of the K components being assigned to a single source through $\mathbf{Q}(\Psi_Q) \equiv 0$ for some appropriate set Ψ_Q of indices, the components of each source are marked as silent through $\mathbf{H}(\Psi_H) \equiv 0$ with an appropriate set Ψ_H of indices.

3 Separation and declipping

Similar to the algorithm introduced in [3], we propose to estimate model parameters using a generalized expectation-maximization (GEM) algorithm [4] and to estimate the signals using the Wiener filtering [6]. The proposed algorithm is briefly described in Algorithm 1, and its steps described below. Note that it can be used not only for joint audio inpainting and source separation, but also for audio inpainting only, setting the number of sources to 1, and for source

[4] $\mathbf{x}^T$ and $\mathbf{x}^H$ represent the non-conjugate transpose and the conjugate transpose of the vector (or matrix) $\mathbf{x}$ respectively.

Algorithm 1 GEM algorithm for NTF model estimation

procedure JOINT-INPAINTING-SSEPARATION-NTF($\mathbf{x}'_c, \Xi', \bar{\mathbf{x}}', \Psi_H, \Psi_Q$)
  Initialize non-negative $\mathbf{Q}, \mathbf{W}, \mathbf{H}$ randomly, set $\mathbf{H}(\Psi_H)$ and $\mathbf{Q}(\Psi_Q)$ to 0
  repeat
    Estimate $\hat{\mathbf{s}}$ (sources), given $\mathbf{Q}, \mathbf{W}, \mathbf{H}, \bar{\mathbf{x}}', \Xi'$ ▷ E-step, see §3.1
    Estimate $\tilde{\mathbf{s}}$ (sources obeying clipping constraint) and $\tilde{\mathbf{P}}$ (posterior power spectra), given $\hat{\mathbf{s}}$, $\mathbf{Q}, \mathbf{W}, \mathbf{H}, \bar{\mathbf{x}}', \Xi'$ and $\mathbf{x}'_c$
    ▷ Applying clipping constraint, see §3.2
    Update $\mathbf{Q}, \mathbf{W}, \mathbf{H}$ given $\tilde{\mathbf{P}}$ ▷ M-step, see §3.3
  until convergence criteria met
end procedure

separation only, when the observed indices of the mixture cover the entire time axis.

3.1 Estimation of sources

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. Hence, Thus, without taking into account the clipping constraint, the source can be estimated in the minimum mean square error (MMSE) sense via Wiener filtering [6] given the covariance tensor $\mathbf{V}$ defined in Section 2 by the model parameters $\mathbf{Q}, \mathbf{W}, \mathbf{H}$.

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

$$\hat{\mathbf{s}}_{jn} = \boldsymbol{\Sigma}^H_{\bar{\mathbf{x}}'_n\mathbf{s}_{jn}} \boldsymbol{\Sigma}^{-1}_{\bar{\mathbf{x}}'_n\bar{\mathbf{x}}'_n} \bar{\mathbf{x}}'_n, \qquad \widehat{\boldsymbol{\Sigma}}_{\mathbf{s}_{jn}\mathbf{s}_{jn}} = \boldsymbol{\Sigma}_{\mathbf{s}_{jn}\mathbf{s}_{jn}} - \boldsymbol{\Sigma}^H_{\bar{\mathbf{x}}'_n\mathbf{s}_{jn}} \boldsymbol{\Sigma}^{-1}_{\bar{\mathbf{x}}'_n\bar{\mathbf{x}}'_n} \boldsymbol{\Sigma}_{\bar{\mathbf{x}}'_n\mathbf{s}_{jn}}, \quad (2)$$

with the definitions $\boldsymbol{\Sigma}_{\mathbf{s}_{jn}\mathbf{s}_{jn}} = \mathrm{diag}([v_{jfn}]_f)$, $\boldsymbol{\Sigma}_{\bar{\mathbf{x}}'_n\mathbf{s}_{jn}} = \mathbf{U}(\Xi'_n)^H \mathrm{diag}([v_{jfn}]_f)$ and $\boldsymbol{\Sigma}_{\bar{\mathbf{x}}'_n\bar{\mathbf{x}}'_n} = \mathbf{U}(\Xi'_n)^H \mathrm{diag}([\sum_j v_{jfn}]_f)\mathbf{U}(\Xi'_n)$ where $\mathbf{U}(\Xi'_n)$ is the $M = F \times |\Xi'_n|$ matrix of columns from $\mathbf{U}$ with index in Ξ'_n.

Note that when there is no noise in the mixture, the resulting estimates for the sources with the wiener filtering will always add up exactly to the observed mixture at the non-clipped support.

3.2 Clipping constraint

For a declipping application, the estimated mixture must have amplitude larger than clipping threshold outside OS in windowed time domain, that is:

$$\mathbf{U}(\{m\})^H \sum_j \hat{\mathbf{s}}_{jn} \times \mathrm{sign}(x'_{mn}) \geq |x'_{mn}|, \ \forall n, \ \forall m \notin \Xi'_n. \quad (3)$$

In order to update the NTF model as described in the following section, the posterior power spectra, $\widetilde{\mathbf{P}} = \left[\tilde{p}_{jfn} = \mathbb{E}\left[\,|s_{jfn}|^2 \middle|\, \bar{\mathbf{x}}'_n; \boldsymbol{\theta}\right]\right]_{j,f,n}$, must be computed. However under the clipping constraint (3), the distribution is no longer Gaussian and the computation of posterior power spectra is no longer computationally simple. Instead we use the *Covariance Projection* method introduced in [3], in which the samples not obeying the constraint (3) after the wiener filtering stage are assumed to be known and equal to the clipping threshold and the wiener filtering step is repeated with rest of the unknowns still assumed to be gaussian distributed. As a result, final estimates of the sources $\tilde{\mathbf{s}}$, which satisfy (3) and the corresponding posterior covariance matrix, $\widetilde{\boldsymbol{\Sigma}}_{\mathbf{s}_{jn}\mathbf{s}_{jn}}$, are obtained. Therefore the posterior power spectra can be computed as

$$\tilde{p}_{jfn} = \mathbb{E}\left[\,|s_{jfn}|^2 \middle|\, \bar{\mathbf{x}}'_n; \boldsymbol{\theta}\right] \cong |\tilde{s}_{jfn}|^2 + \widetilde{\boldsymbol{\Sigma}}_{\mathbf{s}_{jn}\mathbf{s}_{jn}}(f,f). \tag{4}$$

3.3 Updating the model

NTF model parameters can be re-estimated using the multiplicative update (MU) rules minimizing the IS divergence [5] between the the 3-valence tensor of estimated source power spectra $\widetilde{\mathbf{P}}$ and the 3-valence tensor of the NTF model approximation $\mathbf{V}$ defined as $D_{IS}(\widetilde{\mathbf{P}}\|\mathbf{V}) = \sum_{j,f,n} d_{IS}(\tilde{p}_{jfn}\|v_{jfn})$, where $d_{IS}(x\|y) = x/y - \log(x/y) - 1$ is the IS divergence; $\tilde{p}_{jfn}$ is specified by (4) and v_{jfn} is as defined in Section 2. As a result, $\mathbf{Q}, \mathbf{W}, \mathbf{H}$ can be updated with the multiplicative update (MU) rules presented in [8]. These MU rules can be repeated several times to improve the model estimate.

4 Experimental results

In order to observe the performance of declipping and source separation using the proposed algorithm, 5 different music mixtures[5], each composed of 3 sources (bass, drums and vocals), are considered under 3 different clipping conditions. For each mixture with a maximum magnitude of 1 in time domain, 3 clipping levels at the thresholds of 0.2 (heavy clipping), 0.5 (moderate clipping) and 0.8 (light clipping) are considered, resulting in a total of 15 mixtures with different clipping levels. Each mixture is reconstructed by joint declipping and source separation, sequential declipping and source separation and only source separation ignoring the clipping artefacts. The proposed algorithm has been used for all the reconstructions[6] with 15 components ($K = 15$ with 5 components assigned to each source). The STFT is computed using a half-overlapping sine window of 1024 samples (64 ms) and the proposed GEM algorithm is run for 100 iterations. The sources in the mixtures are artificially silenced during a percentage

[5] The mixtures are taken from the professionally produced music recordings of SiSEC 2015 (https://sisec.inria.fr/)

[6] For only declipping, the algorithm is used with a single source, and for only the source separation, the algorithm is used with the observed support set as the entire time axis.

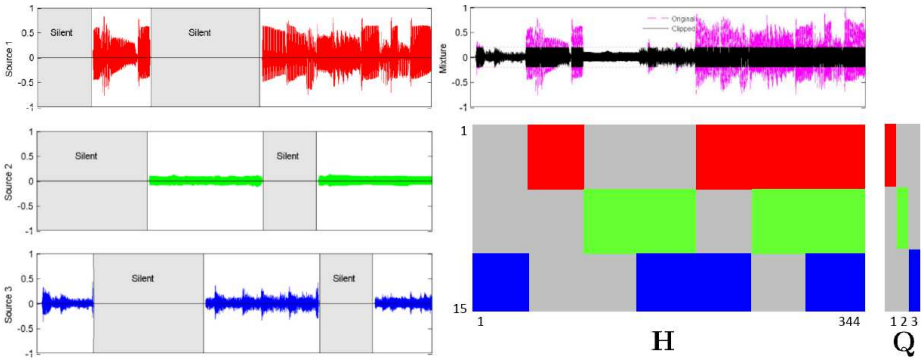

Fig. 1. Typical experimental set-up: a mixture (purple time-domain signal) of 3 sources (red, green, blue) is clipped at 0.2 (black). It will be un-clipped and separated using the low-rank NTF model with each source assigned to 5 out of the $K = 15$ components, as specified by $\mathbf{Q}(\Psi_Q) \equiv 0$ (grey entries), and each source being silent at known instants as specified at the component level by $\mathbf{H}(\Psi_H) \equiv 0$ (grey entries).

of the total time. An example of the activation periods of the sources and corresponding indices set to zero in $\mathbf{Q}$ and $\mathbf{H}$ during reconstruction are shown in Figure 1.

The results of the optimizations can be seen in Table 1. Signal to noise ratio on the clipped support (SNR_m) for the declipped mixture is used to measure the declipping performance and signal to distortion ratio (SDR) is used to measure the source separation performance which are computed as described in [12].

The results in Table 1 show that when the clipping is severe, joint optimization is almost always preferable since it provides improvement on both the quality of the mixture and the quality of the separated sources with respect to source separation without declipping. This is as opposed to sequential approach which provides comparable quality improvement in the mixture at the expense of the performance in source separation. In fact, for heavy clipping the declipping in sequential approach often reduced the performance of source separation noticably with respect to separation without declipping. As the clipping gets lighter, the performance of sequential method approaches to that of joint method, and finally performs slightly better for light clipping. The joint optimization, however, still has few drawbacks which could be improved upon. The declipping in the sequential approach is performed with 15 components without any restrictions whereas the joint optimization is performed with the additional limitation that each source uses 5 components *independently*. Hence it is not possible that two sources share a common component in the joint optimization. This can be overcome by devising better methods to inject the prior information regarding the sources. It should be also noted that the sequential optimization is approximately twice as fast as joint optimization due to handling much less complicated problems in either phases of the sequential processing. The fact that the wiener filtering stage is independent for each window and can be parallelized to provide

Table 1. Performance of joint declipping and source separation ("Joint"), of sequential declipping and source separation ("Sequential") and of source separation only using clipped signal ("S. Separation"), on 5 mixtures of 3 sources and for three levels of clipping from light to heavy. The energy loss percentage due to declipping is also shown in the third column. The declipping performance is measured with SNR_m while the source separation performance is measured with SDR.

	Clipping	**Energy**	**Joint**		**Sequential**		**S. Separation**	
	Loss	**Loss**	SNR_m	SDR	SNR_m	SDR	SNR_m	SDR
	Heavy (th. 0.2)	42.56 %	14.64	9.22	14.14	7.08	7.22	6.01
Mixture 1	Mod. (th. 0.5)	2.60 %	18.78	8.09	19.30	8.10	15.84	8.13
	Light (th. 0.8)	0.04 %	24.49	8.08	25.61	8.07	20.75	8.09
	Heavy (th. 0.2)	50.86 %	9.72	5.13	9.72	5.58	6.62	4.27
Mixture 2	Mod. (th. 0.5)	4.43 %	17.97	6.98	18.57	6.57	14.53	7.11
	Light (th. 0.8)	0.08 %	24.25	6.81	25.15	6.73	21.85	6.76
	Heavy (th. 0.2)	49.28 %	16.64	11.82	17.21	-0.03	7.31	7.79
Mixture 3	Mod. (th. 0.5)	2.52 %	22.41	8.97	22.18	7.07	14.74	9.08
	Light (th. 0.8)	0.09 %	25.45	7.15	24.69	9.12	20.78	9.25
	Heavy (th. 0.2)	50.78 %	7.89	6.11	6.25	4.84	7.44	5.86
Mixture 4	Mod. (th. 0.5)	2.31 %	19.42	9.47	17.88	8.95	15.05	9.14
	Light (th. 0.8)	0.02 %	27.67	9.80	29.45	10.10	19.12	10.14
	Heavy (th. 0.2)	37.11 %	13.60	6.61	13.34	2.76	8.26	5.15
Mixture 5	Mod. (th. 0.5)	1.19 %	18.58	7.85	20.22	8.23	15.20	8.23
	Light (th. 0.8)	0.04 %	17.82	8.10	17.97	8.65	17.73	8.54
	Heavy (th. 0.2)	46.12 %	12.50	7.78	12.13	4.05	7.37	5.82
Average	Mod. (th. 0.5)	2.61 %	19.43	8.27	19.63	7.78	15.07	8.34
	Light (th. 0.8)	0.05 %	23.93	7.99	24.57	8.54	20.05	8.56

significant speed improvements, can be helpful to overcome this problem in the future.

5 Conclusion

Leveraging low-rank NTF techniques, we have proposed a novel framework to attack simultaneously audio inpainting and audio source separation. Focusing on the case where signal loss is due to clipping, we investigated audio declipping and source separation, either jointly or sequentially, with comparison to source separation ignoring the clipping. The results have shown that the clipping artefacts must not be ignored in order to have a good source separation performance, especially in the case of severe clipping. We also showed that, the source separation also improves the performance of declipping and the joint optimization provides better source separation performance in almost all the cases when there is considerable clipping.

The proposed algorithm still has some limitations, such as the reduced flexibility in utilizing the components in the low rank NTF structure. Hence improved

methods to provide prior information on the sources without such limitations is future work. It is also observed that the joint algorithm is slower than the other approaches, and increasing the speed of optimization through better parallel processing is also a promising direction for future research.

References

1. Adler, A., Emiya, V., Jafari, M., Elad, M., Gribonval, R., Plumbley, M.D.: Audio inpainting. IEEE Transactions on Audio, Speech and Language Processing 20(3), 922 – 932 (2012)
2. Bertalmio, M., Sapiro, G., Caselles, V., Ballester, C.: Image inpainting. In: SIGGRAPH'00. pp. 417–424 (2000)
3. Bilen, Ç., Ozerov, A., Pérez, P.: Audio declipping via nonnegative matrix factorization. In: IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (Submitted) (October 2015)
4. Dempster, A., Laird, N., Rubin., D.: Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society. Series B (Methodological) 39, 1–38 (1977)
5. Févotte, C., Bertin, N., Durrieu, J.: Nonnegative matrix factorization with the Itakura-Saito divergence. With application to music analysis. Neural Computation 21(3), 793–830 (Mar 2009)
6. Kay, S.M.: Fundamentals of Statistical Signal Processing: Estimation Theory. Prentice Hall, Englewood Cliffs, NJ (1993)
7. Lee, D., Seung, H.: Learning the parts of objects with nonnegative matrix factorization. Nature 401, 788–791 (1999)
8. Ozerov, A., Févotte, C., Blouet, R., Durrieu, J.L.: 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). pp. 257–260. Prague (May 2011)
9. Roux, J.L., Kameoka, H., Ono, N., de Cheveigné, A., Sagayama, S.: Computational auditory induction as a missing-data model-fitting problem with Bregman divergence. Speech Communication 53(5), 658–676 (May-June 2011)
10. Siedenburg, K., Kowalski, M., Dörfler, M.: Audio declipping with social sparsity. In: Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on. pp. 1577–1581 (May 2014)
11. Simsekli, U., Cemgil, A.T., Yilmaz, Y.K.: Score guided audio restoration via generalised coupled tensor factorisation. In: International Conference on Acoustics Speech and Signal Processing (ICASSP'12). pp. 5369 – 5372 (2012)
12. Vincent, E., Gribonval, R., Févotte, C.: Performance measurement in blind audio source separation. IEEE Trans. Audio, Speech, Language Process. 14(4), 1462–1469 (Jul 2006)
13. Virtanen, T.: Monaural Sound Source Separation by Nonnegative Matrix Factorization With Temporal Continuity and Sparseness Criteria. IEEE Transactions on Audio, Speech and Language Processing 15(3), 1066–1074 (2007)