

NONNEGATIVE MATRIX FACTORIZATION WITH TRANSFORM LEARNING

Dylan Fagot, Herwig Wendt and Cédric Févotte

IRIT, Université de Toulouse, CNRS, Toulouse, France

firstname.lastname@irit.fr

ABSTRACT

Traditional NMF-based signal decomposition relies on the factorization of spectral data, which is typically computed by means of short-time frequency transform. In this paper we propose to relax the choice of a pre-fixed transform and learn a short-time orthogonal transform together with the factorization. To this end, we formulate a regularized optimization problem reminiscent of conventional NMF, yet with the transform as additional unknown parameters, and design a novel block-descent algorithm enabling to find stationary points of this objective function. The proposed joint transform learning and factorization approach is tested for two audio signal processing experiments, illustrating its conceptual and practical benefits.

Index Terms— Nonnegative matrix factorization (NMF), transform learning, single-channel source separation

1. INTRODUCTION

Nonnegative matrix factorization (NMF) has become a privileged approach to spectral decomposition in several fields such as remote sensing and audio signal processing. In the latter field, it has led to state-of-the-art results in source separation [1] or music transcription [2]. The nonnegative data $\mathbf{V} \in \mathbb{R}_+^{M \times N}$ is typically the spectrogram $|\mathbf{X}|$ or $|\mathbf{X}|^{\circ 2}$ of some temporal signal $y \in \mathbb{R}^T$, where \mathbf{X} is a short-time frequency transform of y , $|\cdot|$ denotes the entry-wise absolute value and \circ here denotes entry-wise exponentiation. NMF produces the approximate factorization

$$\mathbf{V} \approx \mathbf{W}\mathbf{H}, \quad (1)$$

where $\mathbf{W} \in \mathbb{R}_+^{M \times K}$ is a nonnegative matrix referred to as *dictionary* that contains spectral patterns characteristic of the data while $\mathbf{H} \in \mathbb{R}_+^{K \times N}$ is the nonnegative matrix that contains the *activation coefficients* that approximate the data samples onto the dictionary. The factorization is usually low-rank ($K < \min(M, N)$) but not necessarily so (in which case regularization constraints should apply on \mathbf{W} and/or \mathbf{H}). The decomposition (1) can then be inverted back to the time domain or post-processed in various ways to solve a large panel of audio signal processing problems. In this traditional setting, a short-time frequency transform such as the short-time Fourier, Cosine or constant- Q transforms acts as a pre-processing of the raw temporal data y . This is a potential limitation as any ill-chosen specification of the time-frequency transform may harm the quality of the decomposition. As such, we here propose to learn the transform *together* with the latent factors \mathbf{W} and \mathbf{H} . We propose to address this task by solving an optimization problem of the form

$$\min_{\phi, \mathbf{W}, \mathbf{H}} D(|\phi(y)|^{\circ 2} | \mathbf{W}\mathbf{H}) \quad (2)$$

This work has received funding from the European Research Council (ERC) under grant agreement No 681839 (project FACTORY). Thanks to Paris Smaragdís for discussions related to this work.

subject to structure constraints on $\phi : \mathbb{R}^T \rightarrow \mathbb{R}^{M \times N}$ and to non-negativity of \mathbf{W} and \mathbf{H} , and where $D(\cdot | \cdot)$ is a measure of fit. In addition, we study and promote the use of sparsity-inducing penalty on \mathbf{H} . We refer to objectives of the form (2) as TL-NMF, which stands for *transform-learning NMF*.

Connections to other works. TL-NMF is inspired by the work of Ravishankar & Bresler [5] on learning sparsifying transforms. Given a collection of data samples \mathbf{Y} (such as a set of images), their work consists in finding an invertible transform Φ such that the output of $\Phi\mathbf{Y}$ is sparse. We are instead looking for a transform ϕ such that $|\phi(y)|^{\circ 2}$ can be well approximated by a NMF. TL-NMF can be viewed as finding a one-layer factorizing network, where y acts as the raw data, ϕ the linear operator, $|\cdot|^{\circ 2}$ the nonlinearity and $\mathbf{W}\mathbf{H}$ the output of the network. Recent work has proposed combining deep learning and NMF, but in a different way. For instance, [6] considers a discriminative NMF setting and [7] study nonnegative auto-encoders. The TL-NMF framework proposed in this work could in addition be extended to fully bridge deep learning and NMF by looking for a cascade of decompositions $f_L(\phi_L \dots f_1(\phi_1(y)))$ such that the output is a NMF. Yet, this is beyond the scope of this paper and left for future work. TL-NMF still operates in a *transformed* domain and is not directly related to synthesis-based NMF models in which the raw data $y(t)$ is modeled as $y(t) = \sum_k c_k(t)$ where the spectrogram of $c_k(t)$ is penalized so as to be closely rank-one [8, 9].

Goals and contributions. The goal of this work is to study the TL-NMF problem of form (2). As a first step, we propose in this paper to gently depart from the traditional short-time Fourier or Cosine transform setting by restricting the transform $\phi(y)$ to be a short-time orthogonal transform (Section 2.1). We consider real-valued transforms for simplicity and use the short-time Cosine transform (STCT) as a baseline. We propose an operational algorithm that returns stationary points of (2) and enables to learn a structured transform ϕ together with the NMF factors \mathbf{W} and \mathbf{H} (Section 2.2). The TL-NMF approach is put to test and compared to conventional STCT-based NMF in two benchmark audio signal processing experiments: music decomposition (Section 3) and speech enhancement (Section 4). The results demonstrate that the proposed approach is operational and yields, in both applications, significant benefits (reduced objective function, data-adapted atoms, improved separation accuracy).

2. NMF MEETS TRANSFORM LEARNING

2.1. Learning a short-time orthogonal transform

Let us denote by $\mathbf{Y} \in \mathbb{R}^{M \times N}$ the matrix that contains adjacent and overlapping short-time frames of size M of y and denote by $\Phi_{\text{DCT}} \in \mathbb{R}^{M \times M}$ the orthogonal real-valued DCT matrix with coefficients $[\Phi_{\text{DCT}}]_{qm} = (2M)^{-1/2} \cos(\pi(q+1/2)(m+1/2)/M)$. With these notations, the power spectrogram $|\mathbf{X}|^{\circ 2}$ of y is simply

Algorithm 1: TL-NMF

Input : $\mathbf{Y}, \tau, K, \lambda$
Output: $\Phi, \mathbf{W}, \mathbf{H}$

Initialize Φ, \mathbf{W} and \mathbf{H}
while $\epsilon > \tau$ **do**

$$\mathbf{H} \leftarrow \mathbf{H} \circ \left[\frac{\mathbf{W}^T ((\mathbf{W}\mathbf{H})^{\circ-2} \circ |\Phi\mathbf{Y}|^{\circ 2})}{\mathbf{W}^T (\mathbf{W}\mathbf{H})^{\circ-1} + \lambda \frac{M}{K} \mathbf{1}_{K \times N}} \right]^{\circ \frac{1}{2}}$$

$$\mathbf{W} \leftarrow \mathbf{W} \circ \left[\frac{((\mathbf{W}\mathbf{H})^{\circ-2} \circ |\Phi\mathbf{Y}|^{\circ 2}) \mathbf{H}^T}{(\mathbf{W}\mathbf{H})^{\circ-1} \mathbf{H}^T + \lambda \frac{M}{K} \mathbf{1}_{M \times N} \mathbf{H}^T} \right]^{\circ \frac{1}{2}}$$

Normalize the columns of \mathbf{W} so that $\|\mathbf{w}_k\|_1 = 1$
Compute γ and Ω as in Section 2.2
 $\Phi \leftarrow \pi(\Phi + \gamma\Omega)$
Normalize Φ to remove sign ambiguity
Compute stopping criterion ϵ as in Eq. (7)

end

given by $|\Phi_{\text{DCT}}\mathbf{Y}|^{\circ 2}$. Traditional NMF (with sparsity) may be cast as

$$\min_{\mathbf{W}, \mathbf{H}} D(|\Phi_{\text{DCT}}\mathbf{Y}|^{\circ 2} | \mathbf{W}\mathbf{H}) + \lambda \frac{M}{K} \|\mathbf{H}\|_1$$

s.t. $\mathbf{H} \geq 0, \mathbf{W} \geq 0, \forall k, \|\mathbf{w}_k\|_1 = 1$ (3)

where the notation $\mathbf{A} \geq 0$ expresses nonnegativity, $\|\mathbf{A}\|_1 = \sum_{ij} |a_{ij}|$ and \mathbf{w}_k denotes the k^{th} column of \mathbf{W} . Regular unpenalized NMF is simply obtained for $\lambda = 0$. The normalizing ratio M/K ensures the measure of fit and the penalty term are of the same order of magnitude. We propose in this work to relax the pre-fixed transform Φ_{DCT} and learn it jointly with \mathbf{W} and \mathbf{H} . In other words, we consider the general case where $\Phi \in \mathbb{R}^{M \times M}$ is a parameter of the objective function defined by

$$C_\lambda(\Phi, \mathbf{W}, \mathbf{H}) \stackrel{\text{def}}{=} D(|\Phi\mathbf{Y}|^{\circ 2} | \mathbf{W}\mathbf{H}) + \lambda \frac{M}{K} \|\mathbf{H}\|_1$$
 (4)

and seek a solution to the TL-NMF problem defined by

$$\min_{\Phi, \mathbf{W}, \mathbf{H}} C_\lambda(\Phi, \mathbf{W}, \mathbf{H})$$

s.t. $\mathbf{H} \geq 0, \mathbf{W} \geq 0, \forall k, \|\mathbf{w}_k\|_1 = 1, \Phi^T \Phi = \mathbf{I}_M$. (5)

We choose at this stage to impose Φ to be orthogonal though one could consider relaxing this assumption as well. The orthogonality constraint implicitly keeps Φ nonsingular and excludes trivial solutions such as $(\Phi, \mathbf{W}, \mathbf{H}) = (\mathbf{0}, \mathbf{0}, \mathbf{0})$ or $(\mathbf{1}_{M \times M}, \mathbf{1}_{M \times 1}, |\mathbf{1}_{1 \times M}\mathbf{Y}|^{\circ 2})$, where $\mathbf{1}_{M \times N}$ denotes the $M \times N$ matrix filled with ones. In this paper, we also choose the measure of fit $D(\cdot | \cdot)$ to be the Itakura-Saito (IS) divergence $D_{\text{IS}}(\mathbf{A} | \mathbf{B}) = \sum_{ij} (a_{ij}/b_{ij} - \log(a_{ij}/b_{ij}) - 1)$. Used with power spectral data, it is known to underlie a variance-structured Gaussian composite model that is relevant to the representation of audio signals [3] and has proven an efficient choice for audio source separation, e.g., [4]. However, the proposed framework can accommodate any other measure of fit.

2.2. Algorithm

We describe a block-coordinate descent algorithm that returns stationary points of problem (5). The blocks are the individual variables \mathbf{W} , \mathbf{H} and Φ that are updated in turn until a convergence criterion is met. As usual of NMF problems, the objective function

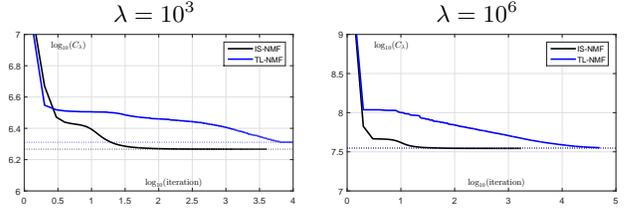


Fig. 1. Values of C_λ w.r.t. iterations for the experiment reported in Section 3.

$C_\lambda(\Phi, \mathbf{W}, \mathbf{H})$ is nonconvex and the returned solution depends on the initialization. We use for \mathbf{W} and \mathbf{H} the standard multiplicative updates that can be derived from a majorization-minimization procedure [10]. The sum-to-one constraint on the columns of \mathbf{W} can be rigorously enforced using a change of variable, like in [11, 12].

Let us now turn our attention towards the update of Φ . We propose to use a gradient-descent procedure with a line-search step selection followed by a projection onto the orthogonal constraint, following the approach of [13]. The main benefit of this approach is that it yields an efficient yet simple algorithm for finding a orthogonal update for Φ . The gradient of the objective function with respect to (w.r.t.) Φ can be shown to be given by

$$\nabla \stackrel{\text{def}}{=} \nabla_{\Phi} C_\lambda(\Phi, \mathbf{W}, \mathbf{H}) = 2(\Delta \circ \mathbf{X})\mathbf{Y}^T$$
 (6)

where $\mathbf{X} = \Phi\mathbf{Y}$, $\Delta = \hat{\mathbf{V}}^{\circ-1} - \mathbf{V}^{\circ-1}$, $\mathbf{V} = |\mathbf{X}|^{\circ 2}$, $\hat{\mathbf{V}} = \mathbf{W}\mathbf{H}$. The steepest manifold-dependent descent direction is given by the natural gradient $\Omega = \Phi \nabla^T \Phi - \nabla$. A suitable step-size γ is then chosen according to the Armijo rule so that the projection $\pi(\Phi + \gamma\Omega)$ of the updated transform onto the orthogonal constraint induces a significant decrease of the objective function [13]. Our block-coordinate descent algorithm is stopped when the relative variation

$$\epsilon^{(i)} = \frac{C_\lambda(\Phi^{(i)}, \mathbf{W}^{(i)}, \mathbf{H}^{(i)}) - C_\lambda(\Phi^{(i-1)}, \mathbf{W}^{(i-1)}, \mathbf{H}^{(i-1)})}{C_\lambda(\Phi^{(i-1)}, \mathbf{W}^{(i-1)}, \mathbf{H}^{(i-1)})}$$
 (7)

between iteration $i - 1$ and i falls below a given threshold τ . The resulting TL-NMF algorithm is summarized in Algorithm 1. The sign ambiguity on Φ resulting from the squared magnitude operation in $|\Phi\mathbf{Y}|^{\circ 2}$ is removed by imposing that the entries of the first column of Φ are positive, likewise the DCT. In the following experiments, we used nonnegative random values for initializing \mathbf{W} and \mathbf{H} . Besides, we chose to initialize the transform Φ in TL-NMF with a random orthogonal matrix.

3. EXPERIMENT 1: MUSIC DECOMPOSITION

In this section, we report results obtained with the proposed algorithm for decomposing real audio data $y(t)$, consisting of a 23 s excerpt of *Mamavatu* by Susheela Raman that has been downsampled to $f_s = 16$ kHz. \mathbf{Y} is constructed using 40 ms-long, 50%-overlapping temporal segments that are windowed with a sine bell. This construction leads to $M = 640$ and $N = 1191$. The proposed TL-NMF is compared to conventional IS-NMF, which is obtained using Alg. 1 as well, but fixed transform $\Phi = \Phi_{\text{DCT}}$. The two algorithms are run with the same stopping threshold $\tau = 10^{-9}$, the same rank $K = 10$, and three values of the regularization parameter

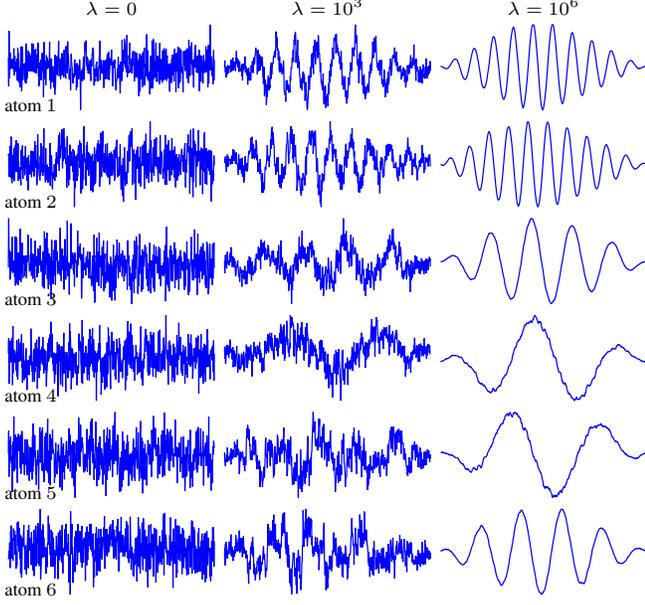


Fig. 2. The six most significant atoms learnt by TL-NMF from random initializations for $\lambda \in \{0, 10^3, 10^6\}$ (from left to right).

$\lambda \in \{0, 10^3, 10^6\}$. The two algorithms are initialized with the same random initializations of \mathbf{W} and \mathbf{H} .

Comparison of objective function values. Fig. 1 displays the objective function values w.r.t. iterations for the two approaches for $\lambda = 10^6$ (the initial objective value is cropped for better readability) and permits the following conclusions. IS-NMF reaches the stopping criterion after fewer iterations than TL-NMF, as expected since the transform Φ is fixed for the latter, leading to faster convergence for the remaining variables \mathbf{W} and \mathbf{H} . Moreover, TL-NMF starts at a larger objective function value due to the random orthogonal transform initialization for Φ . Yet, the proposed TL-NMF algorithm monotonously decreases the objective function to a value that, at convergence, is of the order of the final value of IS-NMF.

Regularization and learnt transform. We now examine examples of the atoms returned by TL-NMF (rows ϕ_m of Φ) for the three values of λ . Fig. 2 displays the six atoms which most contribute to the audio signal (i.e., with largest values $\|\phi_m \mathbf{Y}\|_2$). Clearly, without regularization ($\lambda = 0$), the learnt atoms lack apparent structure. Yet, interestingly, as the value for λ is increased, the atoms become oscillatory and smoother. This is a direct consequence of and justifies the use of the sparsity-inducing term in (3), which induces a structure-seeking constraint on the transform. Eventually, for strong regularization ($\lambda = 10^6$), the learnt atoms resemble packet-like, highly regular oscillations, further analyzed in the next paragraph.

Properties of the learnt atoms. In Fig. 3, pairs of atoms (1, 2), (3, 6), (4, 5) (from left to right) for $\lambda = 10^6$ are plotted, together with the square root of their summed square magnitudes and the sine bell used to window the data. Interestingly, the results demonstrate that the identified pairs of atoms are approximately in *quadrature*. They hence offer shift invariance properties similar to those of the (complex-valued) short time Fourier transform, which is ubiquitous in audio applications. Yet, the TL-NMF algorithm permits to *learn* these atoms from random initializations. This provides strong evidence for the relevance of the proposed approach. Further, note that the atoms embrace the sine bell used to taper the observations \mathbf{Y} .

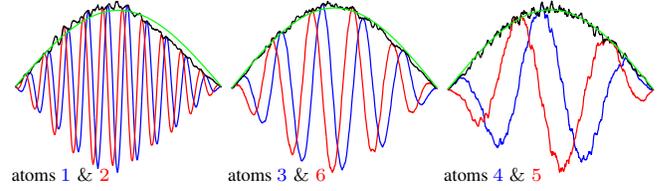


Fig. 3. The atoms in Fig. 2 for $\lambda = 10^6$ plotted pairwise (blue and red solid lines); square root of the sum of their squared magnitude (black solid lines); sine bell window (green solid lines).

4. EXPERIMENT 2: SUPERVISED SOURCE SEPARATION

We now examine whether learning an adaptive transform is actually useful for source separation. To this end, we consider a supervised NMF-based separation setting that follows the approach of [14]. In the following we address the separation of speech from interfering noise, but the method can be applied to any class of sound.

4.1. Principle

We assume that we are given speech and noise reference data $y_{\text{sp}}(t)$ and $y_{\text{no}}(t)$ from which we form short-time matrices \mathbf{Y}_{sp} and \mathbf{Y}_{no} of sizes $M \times N_{\text{sp}}$ and $M \times N_{\text{no}}$, as in Section 2.1. Given a noisy speech recording $y(t)$ with short-time matrix \mathbf{Y} , traditional supervised NMF amounts to estimating activation matrices \mathbf{H}_{sp} and \mathbf{H}_{no} such that

$$\mathbf{V} \approx \mathbf{W}_{\text{sp}} \mathbf{H}_{\text{sp}} + \mathbf{W}_{\text{no}} \mathbf{H}_{\text{no}}, \quad (8)$$

subject to sparsity of \mathbf{H}_{sp} and \mathbf{H}_{no} , where $\mathbf{V} = |\Phi_{\text{DCT}} \mathbf{Y}|^{\circ 2}$, $\mathbf{W}_{\text{sp}} = |\Phi_{\text{DCT}} \mathbf{Y}_{\text{sp}}|^{\circ 2}$, $\mathbf{W}_{\text{no}} = |\Phi_{\text{DCT}} \mathbf{Y}_{\text{no}}|^{\circ 2}$ [14]. Temporal source and noise estimates are then reconstructed in a second step by so-called Wiener filtering [3], based on the spectrogram estimates $\hat{\mathbf{V}}_{\text{sp}} = \mathbf{W}_{\text{sp}} \mathbf{H}_{\text{sp}}$ and $\hat{\mathbf{V}}_{\text{no}} = \mathbf{W}_{\text{no}} \mathbf{H}_{\text{no}}$.

In this section, we generalize this procedure by again learning an optimal transform within the separation procedure. To this end, we propose to build an approximation like (8) but where the fixed transform $\Phi = \Phi_{\text{DCT}}$ is now relaxed and learnt together with \mathbf{H}_{sp} and \mathbf{H}_{no} . This means we propose to minimize

$$\begin{aligned} C_{\Lambda}(\Phi, \mathbf{H}_{\text{sp}}, \mathbf{H}_{\text{no}}) &\stackrel{\text{def}}{=} D_{\text{IS}}(|\Phi \mathbf{Y}|^{\circ 2} \mid |\Phi \mathbf{Y}_{\text{sp}}|^{\circ 2} \mathbf{H}_{\text{sp}} + |\Phi \mathbf{Y}_{\text{no}}|^{\circ 2} \mathbf{H}_{\text{no}}) \\ &+ \lambda_{\text{sp}} \frac{M}{N_{\text{sp}}} \|\mathbf{H}_{\text{sp}}\|_1 + \lambda_{\text{no}} \frac{M}{N_{\text{no}}} \|\mathbf{H}_{\text{no}}\|_1 \\ \text{s.t. } &\Phi^T \Phi = \mathbf{I}, \mathbf{H}_{\text{sp}} \geq 0, \mathbf{H}_{\text{no}} \geq 0, \end{aligned} \quad (9)$$

with $\Lambda = (\lambda_{\text{sp}}, \lambda_{\text{no}})$ defining the possibly different weights in front of the sparsity terms. Note that Φ now appears in both sides of the data-fitting term $D_{\text{IS}}(\cdot \mid \cdot)$, as the same transform is applied to the mixed data \mathbf{Y} and the reference data \mathbf{Y}_{sp} and \mathbf{Y}_{no} . This requires to slightly modify the gradient of C_{Λ} w.r.t. Φ as compared to Section 2 and as described in next section. Given a solution to (9) and $\mathbf{V} = |\Phi \mathbf{Y}|^{\circ 2}$ along with speech and noise spectrogram estimates $\hat{\mathbf{V}}_{\text{sp}} = |\Phi \mathbf{Y}_{\text{sp}}|^{\circ 2} \mathbf{H}_{\text{sp}}$ and $\hat{\mathbf{V}}_{\text{no}} = |\Phi \mathbf{Y}_{\text{no}}|^{\circ 2} \mathbf{H}_{\text{no}}$, temporal estimates may still be produced with Wiener filtering, i.e.,

$$\hat{\mathbf{Y}}_s = \Phi^T \left(\frac{\hat{\mathbf{V}}_{\text{sp}}}{\hat{\mathbf{V}}_{\text{sp}} + \hat{\mathbf{V}}_{\text{no}}} \circ (\Phi \mathbf{Y}) \right) \quad (10)$$

Algorithm 2: Supervised TL-NMF**Input :** $\mathbf{Y}, \mathbf{Y}_{\text{ref}}, \tau, \Lambda = (\lambda_{\text{sp}}, \lambda_{\text{no}})$ **Output:** Φ, \mathbf{H} Initialize Φ, \mathbf{H} **while** $\epsilon > \tau$ **do** $\mathbf{V} = |\Phi \mathbf{Y}|^{\circ 2}, \mathbf{W} = |\Phi \mathbf{Y}_{\text{ref}}|^{\circ 2}$

$$\mathbf{H} \leftarrow \mathbf{H} \circ \left[\frac{\mathbf{W}^T ((\mathbf{W}\mathbf{H})^{\circ -2} \circ \mathbf{V})}{\mathbf{W}^T (\mathbf{W}\mathbf{H})^{\circ -1} + M \left[\frac{\lambda_{\text{sp}}}{N_{\text{sp}}} \mathbf{1}_{N \times N_{\text{sp}}}, \frac{\lambda_{\text{no}}}{N_{\text{no}}} \mathbf{1}_{N \times N_{\text{no}}} \right]^T} \right]^{\circ \frac{1}{2}}$$

Compute γ and Ω as in Section 4.2 $\Phi \leftarrow \pi(\Phi + \gamma \Omega)$ Normalize Φ to remove sign ambiguityCompute stopping criterion ϵ **end**

followed by standard overlap-adding of the columns of $\hat{\mathbf{Y}}_{\text{sp}}$ to return $\hat{y}_{\text{sp}}(t)$, and likewise for the noise. This is exactly the same procedure than in traditional NMF-based separation except that Φ_{DCT} and Φ_{DCT}^T are replaced by Φ and Φ^T .

4.2. Algorithm

Denote $\mathbf{Y}_{\text{ref}} = [\mathbf{Y}_{\text{sp}}, \mathbf{Y}_{\text{no}}]$, $\mathbf{X}_{\text{ref}} = \Phi \mathbf{Y}_{\text{ref}}$, $\mathbf{W} = |\mathbf{X}_{\text{ref}}|^{\circ 2}$, $\mathbf{H} = [\mathbf{H}_{\text{sp}}, \mathbf{H}_{\text{no}}]^T$ and $\hat{\mathbf{V}} = \mathbf{W}\mathbf{H}$. Given \mathbf{W} , \mathbf{H} can be updated with multiplicative rules derived from majorization-minimization as in [10]. We use again a gradient-descent approach for the update of Φ . The gradient of the objective function (9) can be expressed as

$$\nabla_{\Phi} C_{\Lambda}(\Phi, \mathbf{H}) = 2(\Delta \circ \mathbf{X}) \mathbf{Y}^T + 2(\Xi \circ \mathbf{X}_{\text{ref}}) \mathbf{Y}_{\text{ref}}^T \quad (11)$$

where $\Delta = \hat{\mathbf{V}}^{\circ -1} - \mathbf{V}^{\circ -1}$ and $\Xi = \Delta' \mathbf{H}^T$ with $\Delta' = \frac{\hat{\mathbf{V}} - \mathbf{V}}{\hat{\mathbf{V}} \circ \mathbf{V}}$.

Note that the first term of (11) is the gradient in (6). The second term is nothing but the gradient of the data-fitting term D_{IS} with its first argument fixed. Based on (11), we again use a line-search step selection in the steepest natural gradient direction followed by a projection, like in Section 2.2 and following [13]. The resulting algorithm is summarized in Alg. 2.

4.3. Speech enhancement experiment

We consider clean speech and noise data from the TIMIT corpus [15] and the CHIME challenge,¹ respectively. For speech reference data $y_{\text{sp}}(t)$, we use all utterances but the first one in the `train/drl/fc/jf0` directory (about 21 s in total). For noise reference data $y_{\text{no}}(t)$, we use 30 s of the file `BGD_150204_010_BUS.CH1.wav`, which contains noise recorded in a bus. A simulated mixed signal $y(t)$ of duration 3 s is generated by mixing the remaining speech utterance with another segment of the noise file (as such, the test data is not included in the reference data), using signal-to-noise (SNR) ratios of -10 dB and 0 dB. The audio files' sampling frequency is $f_s = 16$ kHz, and short-term matrices \mathbf{Y} , \mathbf{Y}_{sp} and \mathbf{Y}_{no} are constructed using 40 ms-long, 50%-overlapping windowed segments like in Section 3, leading to $M = 640$, $N = 149$, $N_{\text{sp}} = 1059$ and $N_{\text{no}} = 1517$.

Our supervised TL-NMF approach is compared to the traditional supervised NMF procedure (with the IS divergence) described in Section 4.1, based on the same reference data and using the

| Method | SDR (dB) | | SIR (dB) | | SAR (dB) | |
|------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|
| SNR=-10 dB | \hat{y}_{sp} | \hat{y}_{no} | \hat{y}_{sp} | \hat{y}_{no} | \hat{y}_{sp} | \hat{y}_{no} |
| Baseline | -9.50 | 10.00 | -9.50 | 10.00 | ∞ | ∞ |
| IS-NMF | -6.75 | 6.82 | -5.00 | 13.95 | 4.12 | 7.93 |
| TL-NMF | 1.73 | 12.29 | 13.44 | 13.33 | 2.22 | 19.20 |
| SNR=0 dB | \hat{y}_{sp} | \hat{y}_{no} | \hat{y}_{sp} | \hat{y}_{no} | \hat{y}_{sp} | \hat{y}_{no} |
| Baseline | 0.10 | 0.08 | 0.10 | 0.08 | ∞ | ∞ |
| IS-NMF | 1.73 | 0.69 | 3.06 | 5.32 | 9.30 | 3.65 |
| TL-NMF | 6.50 | 5.81 | 12.11 | 9.16 | 8.16 | 9.00 |

Table 1. Source separation performance.

same regularization parameters (only the transform Φ differs between the two approaches), run with the same stopping threshold $\tau = 10^{-7}$. Source separation performance was assessed using the standard BSS_eval criteria [16]. For simplicity we set $\lambda_{\text{sp}} = \lambda_{\text{no}} = \lambda$ and tested various order of magnitudes of λ . The performance of both IS-NMF and TL-NMF were rather robust to the choice of λ , and also offered best individual performance for the same range of values. We report results with the optimal values $\lambda = 10^{-1}$ and $\lambda = 10^{-4}$ for SNR= -10 dB and SNR= 0 dB, respectively. We also compute the performance criteria obtained by $\hat{y}_{\text{sp}} = \hat{y}_{\text{no}} = y/2$ as an indicative baseline. Table 1 reports the comparison results.

The results show that the extra adaptability offered by TL-NMF is clearly beneficial as far as source separation capabilities are concerned. Indeed, TL-NMF dramatically improves the signal to distortion and interference ratios for the speech source by 8.5, 18.5 (SNR=-10 dB) and 4.73, 9.05 (SNR=0 dB) as compared to IS-NMF. However, the signal to artifact ratios are slightly better with IS-NMF, with improvements of -1.9 (SNR=-10 dB) and -1.14 (SNR=0 dB). Audio samples are available online.²

5. CONCLUSION AND FUTURE WORK

We addressed the task of learning the transform underlying NMF-based signal decomposition jointly with the factorization. Specifically, we have proposed a block-coordinate descent algorithm that enables us to find a orthogonal transform Φ jointly with the dictionary \mathbf{W} and the activation matrix \mathbf{H} . To our knowledge, the proposed algorithm is the first operational procedure for learning a transform in the context of NMF. Our preliminary experiments with real audio data indicate that automatically adapting the transform to the signal pays off when seeking latent factors that accurately represent the data. In particular, we were able to retrieve a naturally structured transform which embeds some form of invariance in a first musical example. Then, we achieved source separation performance that compares very favorably against the state-of-the-art in a speech enhancement experiment. Note that although our presentation focused on the processing of audio data, the approach can be adapted to many other settings where NMF is applied to preprocessed data.

Future work will include the estimation of complex-valued transforms which can be directly compared to the STFT, the influence of the initialization of Φ , the influence of the value of K on the learnt transform as well as relaxations of the orthogonal constraint on Φ to mere nonsingular matrices. Also, the use of alternative optimization strategies that lend themselves well to dealing with non-convex problems in high dimension, including stochastic gradient descent, will be investigated.

¹http://spandh.dcs.shef.ac.uk/chime_challenge²https://www.irit.fr/~Cedric.Fevotte/extras/audio_icassp2018.zip

6. REFERENCES

- [1] P. Smaragdis, C. Févotte, G. Mysore, N. Mohammadiha, and M. Hoffman, "Static and dynamic source separation using non-negative factorizations: A unified view," *IEEE Signal Processing Magazine*, vol. 31, no. 3, pp. 66–75, May 2014.
- [2] E. Vincent, N. Bertin, and R. Badeau, "Harmonic and inharmonic nonnegative matrix factorization for polyphonic pitch transcription," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2008.
- [3] 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, 2009.
- [4] B. King, C. Févotte, and P. Smaragdis, "Optimal cost function and magnitude power for NMF-based speech separation and music interpolation," in *Proc. IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, 2012.
- [5] S. Ravishankar and Y. Bresler, "Learning sparsifying transforms," *IEEE Transactions on Signal Processing*, vol. 61, no. 5, pp. 1072–1086, 2013.
- [6] J. L. Roux, J. R. Hershey, and F. Weninger, "Deep NMF for speech separation," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015.
- [7] P. Smaragdis and S. Venkataramani, "A neural network alternative to non-negative audio models," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017.
- [8] C. Févotte and M. Kowalski, "Low-rank time-frequency synthesis," in *Advances in Neural Information Processing Systems (NIPS)*, 2014.
- [9] H. Kameoka, "Multi-resolution signal decomposition with time-domain spectrogram factorization," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015.
- [10] C. Févotte and J. Idier, "Algorithms for nonnegative matrix factorization with the β -divergence," *Neural Computation*, vol. 23, no. 9, pp. 2421–2456, 2011.
- [11] A. Lefèvre, F. Bach, and C. Févotte. Itakura-Saito nonnegative matrix factorization with group sparsity. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2011.
- [12] S. Essid and C. Févotte. Smooth nonnegative matrix factorization for unsupervised audiovisual document structuring. *IEEE Transactions on Multimedia*, vol. 15, no. 2, pp. 415–425, 2013.
- [13] J. H. Manton, "Optimization algorithms exploiting unitary constraints," *IEEE Transactions on Signal Processing*, vol. 50, no. 3, pp. 635–650, 2002.
- [14] P. Smaragdis, B. Raj, and M. V. Shashanka, "Supervised and semi-supervised separation of sounds from single-channel mixtures," in *Proc. International Conference on Independent Component Analysis and Signal Separation (ICA)*, 2007.
- [15] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, and D. S. Pallett, "Timit acoustic-phonetic continuous speech corpus LDC93S1," Philadelphia: Linguistic Data Consortium, Tech. Rep., 1993.
- [16] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1462–1469, 2006.