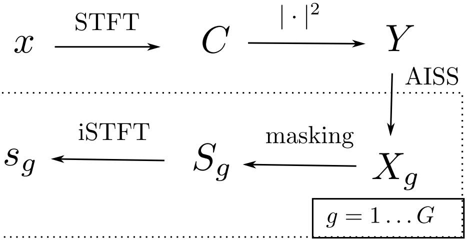
Fast algorithms for informed source separation

Augustin Lefèvre augustin.lefevre@uclouvain.be

July, 10th 2013

Source separation in 5 minutes

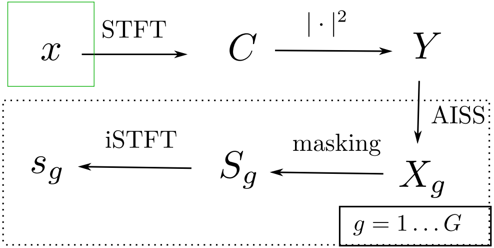


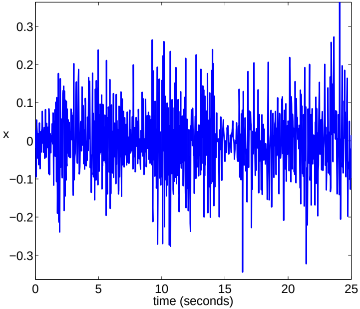
- Recover *source* estimates from a *mixed* signal
- We consider the single-channel setting :

$$x_t = s_t^{(1)} + s_t^{(2)} .$$

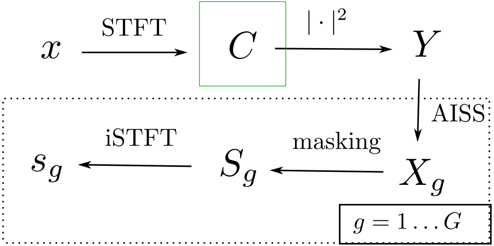
Ill-posed problem, need prior information.

Read mix waveform





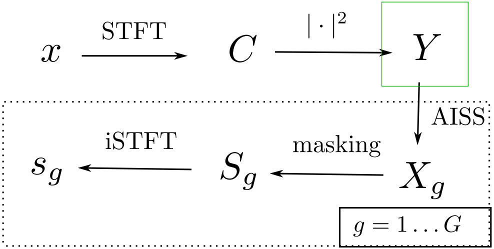
Short time Fourier transform

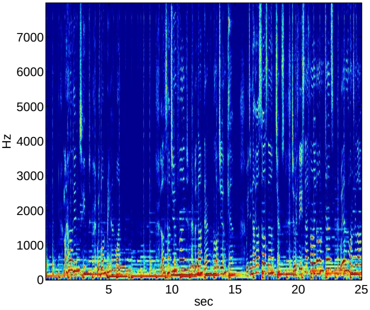


Short time Fourier transform

$$C_{fn} = \sum_{t=1}^{F} x_{t+(n-1)H} w_t \exp\left(-\frac{2(f-1)\pi(t-1)}{F}\right)$$

Remove phase information

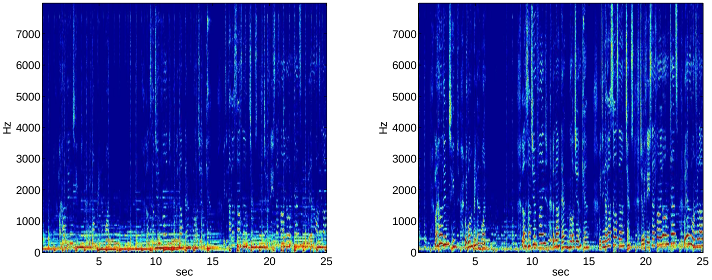




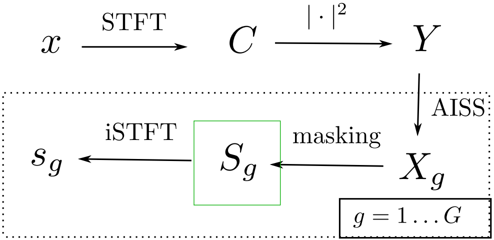
Output of source separation program

$$
\begin{tikzcd}
x \arrow[r, "\mathrm{STFT}"] & C \arrow[r, "|\cdot|^2"] & Y \arrow[d, "\mathrm{AISS}"] \\
s_g & S_g \arrow[l, "\mathrm{iSTFT}"'] & X_g \arrow[l, "\mathrm{masking}"']
\end{tikzcd}
$$

$g = 1 \dots G$



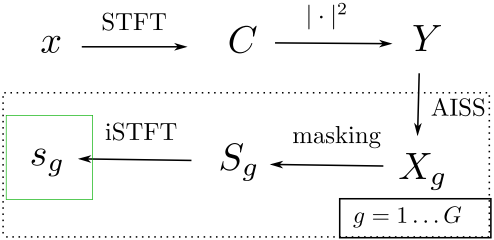
Time-frequency masking

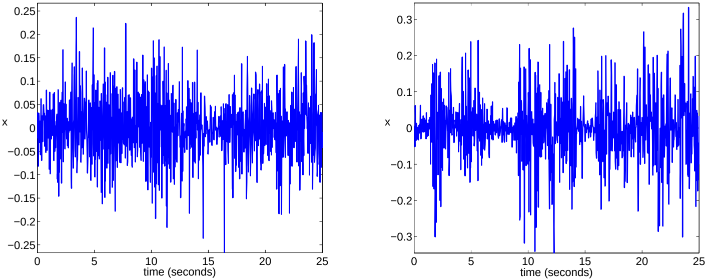


Estimates of each source's complex STFT are obtained by :

$$S_{g,fn} = \frac{X_{g,fn}}{\sum_l X_{l,fn}} C_{fn}$$

Estimate waveforms from STFT





Annotation informed source separation

[Lefèvre et al., 2012, Bryan and Mysore, 2013]: interaction between user and source separation software.

[Lefèvre et al., 2012]: detector trained on development database (random forest, SVM, nearest-neighbour, etc.).

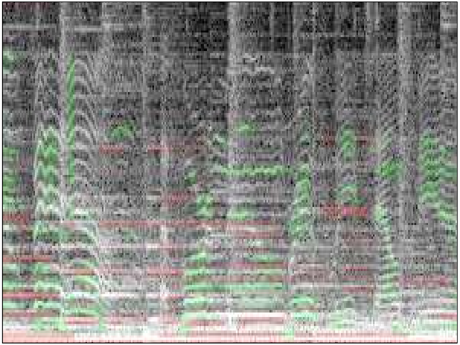


Figure : Detections in the spectrogram

`AISS_nmf`: non-convex

Annotation informed source separation.

Information is used as additional constraints : $M_g \odot X_g = M_g \odot T_g$.

[Lefèvre et al., 2012] : nonnegative matrix factorization (`nmf`) with constraints :

$$\begin{array}{cc} \min_{D,A} & \|Y - \sum_g D_g A_g\|_F^2 \\ \text{s.t.} & D \in \mathbb{R}_+^{F\times K}, A \in \mathbb{R}_+^{K\times N} \\ & M_g \odot (D_g A_g) = M_g \odot T_g \end{array}$$

$Y \in \mathbb{R}_+^{F\times N}$ is the input *spectrogram*.

Need only $D_1 A_1 \geq 0$, but impose stronger constraint : $D_1 \geq 0$, $A_1 \geq 0$ (NMF).

`nmf` is hard ... see talk by Nicolas Gillis.

`AISS_lownuc` : convex

Informed souce separation : $X_1, \dots, X_G \in \mathbb{R}^{F \times N}$.

$$\begin{array}{ll} \min_X & \frac{1}{2}\|Y - \sum_{g=1}^{G} X_g\|_F^2 + \lambda \sum_{g=1}^{G} \|X_g\|_* \\ \text{s.t.} & M_g \odot X_g = M_g \odot T_g \\ & X_g \geq 0 \end{array}$$

The rank of a matrix is revealed in its SVD : $X = P\Sigma Q^\top$.

- $\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_F \geq 0$ singular values.
  $\|X_g\|_* = \sum_{f=1}^{F} \sigma_f$.
  Projecting on $X_g \geq 0$ is straightforward.
  Instead of one `nmf`, we will make repeated calls to `svd` to compute $\|X_g\|_*$ and additional information.

Algorithms for informed source separation

Convex but nonsmooth problem.

Related approaches if no noise and no inequality constraints (Recht et al., 2010) :

$$\begin{array}{ll} \min & \|X\|_* \\ \text{s.t.} & \mathcal{A}(X) = b \end{array}$$

where $\mathcal{A} : \mathbb{R}^{F \times NG} \mapsto \mathbb{R}^p$, $b \in \mathbb{R}^p$ $(p \ll m \times n)$ is linear.

Link with SDP optimization :

$$\begin{array}{ll} \min & t \\ \text{s.t.} & \mathcal{A}(X) = b \\ & \begin{pmatrix} tI & X \\ X & tI \end{pmatrix} \succeq 0 \end{array}$$

Use interior-point solver, which has superlinear convergence rate.

BUT Hessian has size $O(F^2 N^2)$, i.e. 10^{10} for a ten seconds audio track. This is too large !

Subgradient descent

Objective function f is convex so it admits derivatives in all directions :

$$f'(X; D) = \lim_{t \downarrow 0} \frac{f(X + tD) - f(X)}{t}$$

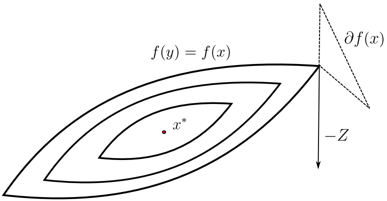
Subgradients generalize the gradient :

$$Z \in \partial f(X) \leftrightarrow f'(X; D) \geq \langle Z, D \rangle$$

$\langle Z, D \rangle = \sum_g \mathrm{Tr}\ Z_g^\top D_g$

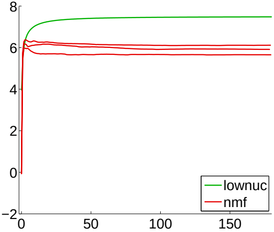
Projected subgradient descent : $X^{(t+1)} = \Pi(X^{(t)} - \mu_t Z^{(t)})$.

Warning : $f(X^{(t+1)}) \not\leq f(X^{(t)})$.

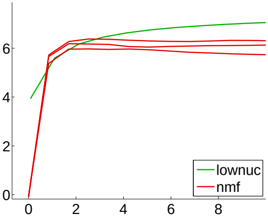


Guarantee :$\mu_t = \mu_0 (1 + t)^{-\frac{1}{2}} \Rightarrow \|X^{(t)} - X^*\| \searrow 0$.

Controlled experiments



(a) Overall



(b) First few seconds

Figure : (Left) Evolution of SDR as a function of CPU time (in seconds), for (green) our method and (red) NMF started from several initial points.

SDR is a measure of how well we have separated sources (the higher the better).

Shrinkage of singular values

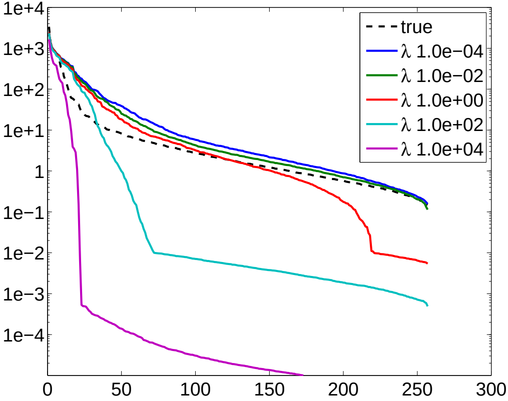


Figure : Magnitude of singular values in decreasing order, for various values of λ. Dotted line is the true singular value profile.

Smoothing technique [Nesterov, 2003]

$$\begin{array}{ll} \min_X & \frac{1}{2}\|Y - \sum_{g=1}^{G} X_g\|_F^2 + \lambda \sum_{g=1}^{G} \|X_g\|_{*,\mu} \\ \text{s.t.} & M_g \odot X_g = M_g \odot T_g \\ & X_g \geq 0 \end{array}$$

$\|\cdot\|_{*,\mu}$ is $C^{(1,1}$ with Lipschitz constant $\frac{1}{\mu}$ and :

$$\|X\|_{*,\mu} \leq \|X\|_* \leq \|X\|_{*,\mu} + \mu C \qquad \forall X \in \mathbb{R}^{F \times N}$$

$$\|X\|_* = \max\{\mathrm{Tr}\ U^\top X,\ \sigma_1(U) \leq 1\}$$

$$\|X\|_{*,\mu} = \max\{\mathrm{Tr}\ U^\top X - \|U\|_F^2,\ \sigma_1(U) \leq 1\}$$

Apply accelerated gradient descent to the smooth minimization problem.

$\mu = 0$: slow convergence but accurate solutions.

Large μ : fast but inaccurate solutions.

Comparison with subgradient

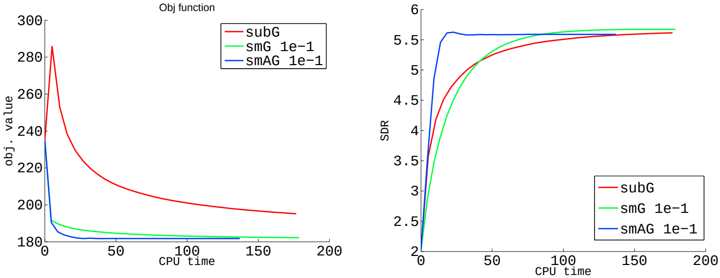


Figure : Decrease of the objective function as a function of the allowed CPU time, for various algorithms

Effect of μ

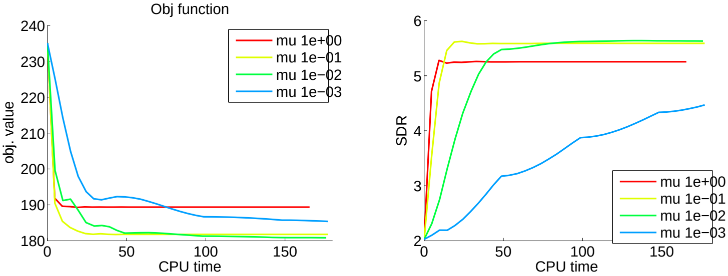


Figure : Decrease of the objective function as a function of the allowed CPU time, for various values of μ.

We display the *original* objective function :

$$\frac{1}{2}\| Y - \sum_{g=1}^{G} X_g \|_F^2 + \lambda \sum_{g=1}^{G} \|X_g\|_* \, .$$

Conclusion

Our formulation contributes to the field of *informed* source separation methods, where knowledge is directly *relevant* to the query audio track, and involves *interaction with the user*.

These methods are the state of the art in single-channel source separation benchmarks.

Our convex formulation compares well with its NMF counterpart, even with a subgradient algorithm.

The smoothing technique allows to retrieve more accurate solutions for a given CPU budget.

More complex constraints ? E.g., source estimates should classify correctly : $\langle W, X_g \rangle + b \leq 0$.

Proximal operator :

$$\operatorname{prox}(\bar{X}) = \begin{array}{cc} \arg\ \min_X & \frac{1}{2}\|\bar{X} - X\|_F^2 + \lambda\|X\|_* \,, \\ \text{s.t.} & M_g \odot X_g = M_g \odot T_g \,, \end{array}$$

Necessary and sufficient conditions :

$$\begin{aligned}
&0 \in X - \bar{X} + \lambda(PQ^\top + W) + M \odot E \\
&W^\top X = 0 \\
&WX^\top = 0 \\
&M \odot X = M \odot T \\
&\|W\|_{\text{op}} \leq 1
\end{aligned}$$

where $E \in \mathbb{R}^{F \times N}$ are Lagrangian multiplicators associated with the constraint $M \odot X = 0$. Note that here, $X = P\Sigma Q^\top$ is an economy-size SVD of X *and not* $\bar{X}$, so P and Q depend on X.

N.J. Bryan and G.J. Mysore. Interactive Refinement of Supervised and Semi-supervised Sound Source Separation Estimates. In *ICASSP*, 2013.

J.-L. Durrieu, B. David, and G. Richard. A Musically Motivated Mid-Level Representation For Pitch Estimation And Musical Audio Source Separation. *IEEE Journal of Selected Topics on Signal Processing*, 5(6):1180–1191, Oct. 2011.

A. Lefèvre, F. Bach, and C. Févotte. Semi-supervised NMF with time-frequency annotations for single-channel source separation. In *International Conference on Music Information Retrieval (ISMIR)*, 2012.

Y. Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer, 2003.