SPARSE DECOMPOSITION OF STEREO SIGNALS WITH MATCHING PURSUIT AND APPLICATION TO BLIND SEPARATION OF MORE THAN TWO SOURCES FROM A STEREO MIXTURE

R. Gribonval

METISS Project
IRISA-INRIA
Campus de Beaulieu, F-35042 Rennes Cedex, France
remi.gribonval@inria.fr

ABSTRACT

We develop a method of sparse decomposition of stereo audio signals, and test its application to blind separation of more than two sources from only two linear mixtures. The decomposition is done in a *stereo dictionary* which we can define based on any standard time-frequency or time-scale dictionary, such as the multiscale Gabor dictionary. A decomposition of a stereo mixture in the dictionary is computed with a Matching Pursuit type algorithm called *Stereo Matching Pursuit*. We experiment an application to blind source separation with three (mono) sources mixed on two channels. We cluster the parameters of the stereo atoms of the decomposition to estimate the mixing parameters, and recover estimates of the sources by a partial reconstruction using only the appropriate atoms of the decomposition. The method outperforms the best achievable linear demixing by 3 dB to more than 7 dB on our preliminary experiments, and its performance should increase as we let the number of iterations of the pursuit increase. Sample sound files can be found here : http://www.irisa.fr/metiss/gribonval/

1. INTRODUCTION

Stereo audio signals can be modeled as a pair of (noisy) mixtures

$$y_c(t) = \sum_{i=1}^{I} (h_{c,i} \star x_i)(t) + n_c(t), \qquad c = l, r \quad (1)$$

of $I \geq 2$ (mono) sources $x_1(t), \ldots, x_I(t)$, with an additive (stereo) Gaussian noise $n(t) := (n_l(t), n_r(t))$ and $h_i(t) := (h_{l,i}(t), h_{r,i}(t))$ a pair of linear filters.

In the natural acoustic mixing that occurs during the simultaneous recording of several sources with a pair of microphones, each pair of filters $h_i(t)$ depends on the spatial location of the source relatively to the sensors. In the anechoic case, they can be modeled as gain-delay filters, and the stereo mixture $y(t) = (y_l(t), y_r(t))$ can be expressed as

$$\sum_{i=1}^{I} \lambda_i \Big(\cos\Theta_i \;\, x_i(t-\tau_{l,i}), \sin\Theta_i \;\, x_i(t-\tau_{r,i}) \Big) + n(t) \quad (2)$$

where Θ_i is a *panpot parameter* : $\Theta_i = 0$ corresponds to mixing x_i entirely on the left channel, $\Theta_i = \pi/2$ mixes it to the right channel.

Mono audio sources $x_i(t)$, which can be considered as vectors in the Hilbert space $\mathcal{H}$ of finite energy signals, have been shown to have sparse decompositions in a variety of time-frequency dictionaries (*e.g.* local trigonometric bases [1], wavelet or wavepacket bases [2] or the union of them [3], or the Gabor multiscale dictionary [4, 5, 6]). By sparse representation we mean that $x_i = \sum_k a_{i,k} g_k$ with $g_k \in \mathcal{D}$ (we denote the dictionary by $\mathcal{D}$) and the sequence $\{a_{i,k}\}$ has a fast decay when k tends to infinity. It follows that the stereo mixture y, which lies in the Hilbert space $\mathcal{H}_{\text{stereo}} := \mathcal{H} \oplus \mathcal{H}$ of finite energy stereo signals, has a representation as

$$\sum_{i=1}^{I} \sum_k a_{i,k} \lambda_i \Big(\cos\Theta_i \;\, g_k(t-\tau_{l,i}), \sin\Theta_i \;\, g_k(t-\tau_{r,i}) \Big) + n(t) \quad (3)$$

in the stereo dictionary $\mathcal{D}_{\text{stereo}}$ of stereo atoms

$$(\cos\theta \;\, g(t), \sin\theta \;\, g(t-\tau)), \quad (4)$$

where $g \in \mathcal{D}$ is a (mono) atom, θ is a panpot parameter, and $\tau \in \mathbb{R}$ a delay parameter which can be restricted to $|\tau| \leq \tau_{\max}$ where $\tau_{\max}$ is the maximum delay between channels.

Given the stereo mixture y, we propose to decompose it on a stereo dictionary using a Matching Pursuit type algorithm [5]. After M iterations, $y(t)$ is decomposed as

$$\sum_{m=1}^{M} \alpha_m (\cos\theta_m \;\, g_m(t), \sin\theta_m \;\, g_m(t-\tau_m)) + R^M(t) \quad (5)$$

where $R^M(t) = (R_l^M, R_r^M)(t)$ is a residual. We define the Stereo Matching Pursuit algorithm in Section 2.

In Section 3 we experiment a simple blind source separation algorithm based on Stereo Matching Pursuit decomposition. The idea, which was exploited in [7] using the complex spectrogram as a sparse representation of each channel, is that each stereo atom in the decomposition (5) corresponds to an atom in the representation (3) for some i, hence every pair (θ_m, τ_m) from (5) yields an estimate of the panpot and relative delay $(\Theta_i, \tau_{r,i} - \tau_{l,i})$. By clustering (θ_m, τ_m) one can estimate the number $\widehat{I}$ of sources and partition the indexes $1 \leq m \leq M$ into $\widehat{I}$ classes $K_i := \{m \, : m \in K_i\}$. As mentioned in [7], in the case of $I = 2$ sources, it is possible to estimate a demixing matrix from the clusters, and the sources can be estimated linearly by applying the demixing matrix. We are more interested in the case of $I > 2$ sources : then, no linear demixing can in general completely separate the sources. However by partitioning the decomposition (5), one can obtain a nonlinear estimate of the sources (up to gain and delay)

$$\widehat{\lambda_i x_i} := \sum_{m \in K_i} \alpha_m g_m. \tag{6}$$

In a way, this extends the complex spectrogram-based blind source separation technique proposed in [7] by providing a method for adaptively choosing the size of the window.

2. STEREO MATCHING PURSUIT

We recall the definition of the Matching Pursuit algorithm [5] and specialize it to the setting of stereo signals. Given a complete dictionary $\mathcal{D}$, *i.e.* a family of unit vectors in a Hilbert space $\mathcal{H}$ that spans a dense subspace of $\mathcal{H}$ (note : it can easily be checked that if $\mathcal{D}$ is complete in $\mathcal{H}$ then $\mathcal{D}_{\text{stereo}}$ is complete in $\mathcal{H}_{\text{stereo}}$), and an arbitrary number M, the Matching Pursuit decomposes any signal $y(t)$ into a linear combination of M atoms chosen among $\mathcal{D}$ and a residual term $R^M(t)$ as in (5). The strong convergence of the algorithm $\lim_{M\to\infty} \|R^M\| = 0$ was proved by Jones [8] and shows that one can get as good an approximation to $y(t)$ as wanted by taking M big enough.

2.1. Standard algorithm

Standard Matching Pursuit goes as follows. From a decomposition of y into $M - 1 \geq 0$ atoms, one gets an M-atom decomposition in the following way :

1. Compute $\left|\left\langle R^{M-1}, g\right\rangle\right|$ for all $g \in \mathcal{D}$.
2. Select the best atom of the dictionary
$$g_M := \arg\max_{g\in\mathcal{D}} \left|\left\langle R^{M-1}, g\right\rangle\right|.$$
3. Compute the new residual
$$R^M(t) := R^{M-1}(t) - \alpha_M g_M(t) \tag{7}$$
with $\alpha_M := \left\langle R^{M-1}, g_M\right\rangle$.

2.2. Stereo Matching Pursuit

For any mono atom $g \in \mathcal{D}$ and delay parameter τ, the pair of stereo vectors $\{(g, 0), (0, g(t-\tau))\}$ is an orthonormal basis of its linear span $\mathcal{V}_{g,\tau}$ in $\mathcal{H}_{\text{stereo}}$, which we will call a *stereo subspace*. The orthonormal projection $P_{\mathcal{V}_{g,\tau}} R^{M-1}$ of R^{M-1} onto $\mathcal{V}_{g,\tau}$ is given by

$$\Big(\left\langle R_l^{M-1}, g\right\rangle g, \left\langle R_r^{M-1}(t), g(t-\tau)\right\rangle g(t-\tau)\Big) \tag{8}$$

and one can easily check that

$$\begin{aligned}\max_{h\in\mathcal{V}_{g,\tau},\|h\|=1} \left|\left\langle R^{M-1}, g\right\rangle\right|^2 &= \|P_{\mathcal{V}_{g,\tau}} R^{M-1}\|^2 \\ &= \left|\left\langle R_l^{M-1}, g\right\rangle\right|^2 + \left|\left\langle R_r^{M-1}(t), g(t-\tau)\right\rangle\right|^2.\end{aligned}$$

Assume that for any $g \in \mathcal{D}$ and τ, $g(t-\tau) \in \mathcal{D}$: it follows that Matching Pursuit with $\mathcal{D}_{\text{stereo}}$ goes as :

1. Compute $\left\langle R_c^{M-1}, g\right\rangle$ for $c = l, r$ and all $g \in \mathcal{D}$.
2. Compute $\|P_{\mathcal{V}_{g,\tau}} R^{M-1}\|^2$ for all $g \in \mathcal{D}$ and τ.
3. Select the best stereo subspace $\mathcal{V}_{g_M,\tau_M}$
$$(g_M, \tau_M) := \arg\max_{g\in\mathcal{D}, |\tau|\leq\tau_{\max}} \|P_{\mathcal{V}_{g,\tau}} R^{M-1}\|$$
4. Compute the new residual
$$\begin{aligned} R^M(t) &:= R^{M-1}(t) - \\ &\quad \alpha_M\Big(\cos\theta_M \;\; g_M(t), \sin\theta_M \;\; g_M(t-\tau)\Big)\end{aligned}$$
with $\alpha_M := \|P_{\mathcal{V}_{g_M,\tau_M}} R^{M-1}\|$ and
$$e^{j\theta_M} := \frac{\left\langle R_l^{M-1}, g\right\rangle + j\left\langle R_r^{M-1}(t), g(t-\tau)\right\rangle}{\alpha_M}. \tag{9}$$

Hence, no exhaustive search over the panpot parameter θ is needed for the optimization of a stereo atom. The complexity of M iterations of Stereo Matching Pursuit for a signal of N samples is essentially twice that of standard Matching Pursuit, *i.e.* $\mathcal{O}(MN\log^2 N)$ [5, 9] with the usual discretization of the Gabor dictionary [4].

3. EXPERIMENTS

Using the Matching Pursuit Package of the LastWave program [10] we have implemented Stereo Matching Pursuit using the stereo dictionary of real Gabor atoms [4, 5]

$$g_{s,u,\xi,\phi}(t) := c_{s,\xi,\phi} w \left(\frac{t-u}{s}\right) \cos\left(2\pi\xi(t-u)+\phi\right), \tag{10}$$

where $w(t)$ is a given window of unit energy, *e.g.* the Gaussian window, s is a scale parameter, u a time parameter, ξ a frequency parameter and $c_{s,\xi,\phi}$ a normalizing constant. Each real Gabor atom is in the linear span of two conjugated complex Gabor atoms

$$g(t) = g_{s,u,\pm\xi}(t) := \frac{1}{\sqrt{s}} w \left(\frac{t-u}{s}\right) e^{\pm j2\pi\xi(t-u)} \tag{11}$$

therefore, similarly to what was done above with the panpot parameter θ_M, no exhaustive search over the phase parameter ϕ_M is needed [11, 12, 9].

3.1. Stereo Matching Pursuit of real audio signals

We performed experiments on a mixture of three sources : $x_1(t)$ is a recording of cello; $x_2(t)$ is a recording of drums; $x_3(t)$ is a recording of piano. Each source is sampled at 8 kHz and we use 2.4 seconds of each signal, *i.e.* $N = 19200$ samples. Stereo Matching Pursuit was performed using $M = 2000$ iterations. The computation time on a Pentium III 750 MHz laptop was about 30 minutes. We display on Figure 1 the decay (in decibels) of the relative error $\|R^M\|/\|y\|^2$ as a function of M. Due to the relatively high

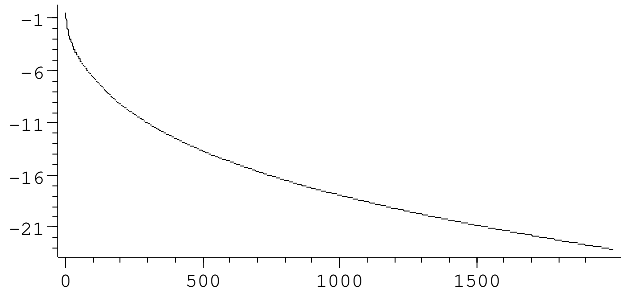

Fig. 1. Decay (in decibels) of the relative approximation error $\|R^M\|/\|y\|^2$ as a function of the number M of stereo atoms.

computation time of this straightforward implementation of the decomposition algorithm, we have performed only few experiments at the time of writing this paper. However, a fast version of the Matching Pursuit decomposition is under development, based on sub-dictionaries of local maxima of the stereo Gabor dictionary [9]. We expect the computation time to be divided by about 25, thus enabling more iterations and better approximations.

3.2. Source separation of panpot mixture

We tested the source separation capabilities of our decomposition method in the case where the sources are mixed using pure panpot [13], *i.e.* (see Equation (2)) $\tau_{l,i} = \tau_{r,i} = 0$, $\lambda_1 = 1$, $\Theta_1 \approx 0.39$. $\lambda_2 = 1$, $\Theta_2 =\approx 0.79$. $\lambda_3 = 2$, $\Theta_1 \approx 1.18$ and $n(t) = 0$. For each value of Θ_i the difference in intensity between channels $|20 \log \tan \Theta_i|$ is at most 7.7 decibels, hence each of the three sources is perceived strictly between the two ears in binaural hearing [14].

Stereo Matching Pursuit was performed with $\tau_{\max} = 0$. Figure 2 displays the histogram of $\{\theta_m\}_{m=1}^{M}$: three peaks can clearly be observed, they are centered on the values $\Theta_1, \Theta_2, \Theta_3$. Estimates of the three sources where computed

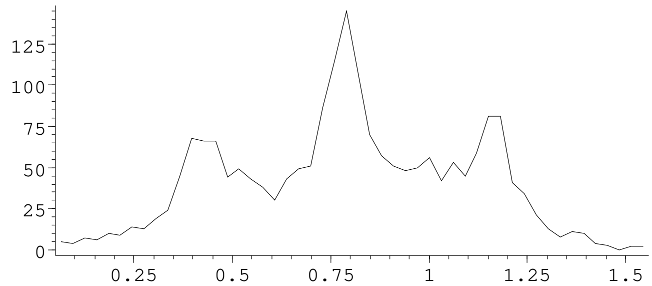

Fig. 2. Histogram of θ_m from a Stereo Matching Pursuit decomposition of a panpot mixture of three audio signals. One can observe three peaks, centered around the values $\Theta_1 \approx 0.39, \Theta_2 \approx 0.79$ and $\Theta_3 \approx 1.18$ (see text).

from Equation (6) by manual clustering in three intervals $\theta \in [0, 0.6], \theta \in (0.6, 1)$ and $\theta \in [1, \pi/2]$. For each source,

source	cello	drums	piano
$\|\lambda_i x_i\|$	2000	4000	5500
$\sharp K_i$	536	924	540
SNR^{mp} (dB)	4.8	8.5	14.6
SNR^{lin} (dB)	0.0	1.1	11.5

Table 1. Signal to noise ratio between the original and the estimated sources with the Stereo Matching Pursuit based separation (SNR^{mp}) and the best linear demixing (SNR^{lin}), and their correlation with the contribution $\|\lambda_i x_i\|$ of the source to the stereo mixture and the number $\sharp K_i$ of atoms used in the estimation (see Equation (6)).

we summarized in Table 1, the norm $\|\lambda_i x_i\|$ of its contribution to the mixture, the number $\sharp K_i$ of atoms used in its estimation, the absolute signal to noise ratio (SNR) in decibels $\mathrm{SNR}^{mp} := 10\log_{10}\|x_i\|^2/\|x_i - \widehat{x}_i\|^2$ and, as a reference, the value SNR^{lin} of the best SNR attainable by linear demixing. One can notice that the SNR logically increases as the contribution of the source to the stereo mixture increases. Moreover, Stereo Matching Pursuit separation outperforms the best linear demixing by 3 dB (for the piano) to more than 7 dB (for the drums). It seems that the larger the number $\sharp K_i$ of atoms used in the estimation of a source, the larger the improvement over linear demixing. Hence, we expect the SNR to improve if we make more iterations of the Stereo Matching Pursuit.

4. CONCLUSION AND ONGOING WORK

In this paper we presented a new method of decomposition of stereophonic audio signals, using the notion of stereo time-frequency dictionary and a Matching Pursuit approach. With the stereo Gabor dictionary, we performed blind source separation by clustering the decomposition coefficients and partial reconstruction. Among other potential applications of the decomposition method, let us point out the modification of the stereo image (*i.e.* remixing at the user end) by changing the θ and τ parameters before reconstruction, as well as techniques of compression of audio signals where, depending on the available bitrate, we may choose to respect more or less the stereo image by using more or less bits to code θ and τ.

At the time of writing this paper, we were about to test an implementation of a fast version of the Stereo Matching Pursuit decomposition, which we expect to multiply the computation speed by 25. Using the fast algorithm, we will make experiments with more iterations of the pursuit, leading to smaller energy of the residue R^M. Therefore, we expect to see an improvement in the SNR of the blind separation application.

Because acoustic stereo recordings generally involve a combination of phase and intensity stereophony, we plan to turn to source separation with nonzero delays $\tau_{c,i}$. Hence, we will have to deal with the fact that the relative delay is more reliably estimated from short atoms of the decomposition (*i.e.* with a small scale s), while the phase difference is more reliable for longer atoms. Moreover, we will likely need to cluster simultaneously the panpot, delay and phase difference parameters $(\theta_m, \tau_m, \phi_{l,m} - \phi_{r,m})$ of the decomposition atoms. We may eventually have to investigate modified decomposition algorithms where some automatic clustering is done adaptively at each iteration of the pursuit, driving the selection of the next atom.

5. REFERENCES

[1] J. Berger, R. Coifman, and M.J. Goldberg, "A method of denoising and reconstructing audio signals," in *Proc. Int. Computer Music Conf. (ICMC'94)*, Sept. 1994, pp. 344–347.

[2] S. Mallat, *A Wavelet Tour of Signal Processing*, Academic Press, 1998.

[3] L. Daudet, *Représentations structurelles de signaux audiophoniques : méthodes hybrides pour des applications à la compression*, Ph.D. thesis, Université de Provence (Aix-Marseille I), 2000.

[4] B. Torrésani, "Wavelets associated with representations of the affine Weyl-Heisenberg group," *J. Math. Phys.*, vol. 32, pp. 1273–1279, May 1991.

[5] S. Mallat and Z. Zhang, "Matching pursuit with time-frequency dictionaries," *IEEE Trans. Signal Process.*, vol. 41, no. 12, pp. 3397–3415, Dec. 1993.

[6] S. Qian and D. Chen, "Signal representation using adaptive normalized Gaussian functions," *Signal Process.*, vol. 36, no. 1, pp. 1–11, 1994.

[7] A. Jourjine, S. Rickard, and O. Yilmaz, "Blind separation of disjoint orthogonal signals: Demixing n sources from 2 mixtures," in *Proc. Int. Conf. Acoust. Speech Signal Process. (ICASSP'00)*, Istanbul, Turkey, June 2000, vol. 5, pp. 2985–2988.

[8] L.K. Jones, "On a conjecture of Huber concerning the convergence of PP-regression," *The Annals of Statistics*, vol. 15, pp. 880–882, 1987.

[9] R. Gribonval, "Fast matching pursuit with a multiscale dictionary of Gaussian chirps," *IEEE Trans. Signal Process.*, vol. 49, no. 5, pp. 994–1001, May 2001.

[10] E. Bacry, *LastWave software (GPL license)*, http://wave.cmap.polytechnique.fr/soft/LastWave/.

[11] F. Bergeaud, *Représentations adaptatives d'images numériques,* Matching Pursuit, Ph.D. thesis, Ecole Centrale Paris, 1995.

[12] M. Goodwin, "Matching pursuit with damped sinusoids," in *Proc. Int. Conf. Acoust. Speech Signal Process. (ICASSP'97)*, 1997.

[13] M. Van Hulle, "Clustering approach to square and non-square blind source separation," in *IEEE Workshop on Neural Networks for Signal Processing (NNSP99)*, Aug. 1999, pp. 315–323.

[14] C. Hugonnet and P. Walder, *Théorie et pratique de la prise de son stéréophonique*, Eyrolles, Paris, 1994.