Extension of sparse, adaptive signal decompositions to semi-blind audio source separation

Andrew Nesbit, Emmanuel Vincent, Mark Plumbley

Extension of Sparse, Adaptive Signal Decompositions to Semi-Blind Audio Source Separation

Andrew Nesbit[1⋆], Emmanuel Vincent[2], and Mark D. Plumbley[1⋆⋆]

[1] Queen Mary University of London
School of Electronic Engineering and Computer Science
Mile End Road, London, E1 4NS, United Kingdom
{andrew.nesbit,mark.plumbley}@elec.qmul.ac.uk
[2] METISS Group, IRISA-INRIA
Campus de Beaulieu, 35042 Rennes Cedex, France
emmanuel.vincent@irisa.fr

Abstract. We apply sparse, fast and flexible adaptive lapped orthogonal transforms to underdetermined audio source separation using the time-frequency masking framework. This normally requires the sources to overlap as little as possible in the time-frequency plane.
In this work, we apply our adaptive transform schemes to the semi-blind case, in which the mixing system is already known, but the sources are unknown. By assuming that exactly two sources are active at each time-frequency index, we determine both the adaptive transforms and the estimated source coefficients using ℓ^1 norm minimisation. We show average performance of 12–13 dB SDR on speech and music mixtures, and show that the adaptive transform scheme offers improvements in the order of several tenths of a dB over transforms with constant block length. Comparison with previously studied upper bounds suggests that the potential for future improvements is significant.

1 Introduction

Our goal is to tackle the problem of *audio source separation* for *underdetermined* and *instantaneous* mixtures. Specifically, given an observed two-channel mixture $x(n) = (x_1(n), x_2(n))$, we aim to estimate all $J > 2$ simultaneously active sources $s(n) = (s_1(n), \ldots, s_J(n))$, assuming the mixture has been generated thus:

$$x(n) = As(n) \ , \tag{1}$$

where $A = (a_{i,j})$ is a $2 \times J$ matrix with real-valued entries $a_{i,j}$, the mixture and source indices are i and j respectively, and the discrete-time index ranges as $0 \leq n < N$.

⋆ AN is supported by EPSRC Grant EP/E045235/1.
⋆⋆ MDP is supported by EPSRC Leadership Fellowship EP/G007144/1.

In the *blind* case only $x(n)$ is known. If $s(n)$ remains unknown but A is given, then the problem is called *semi-blind*. If both A and $s(n)$ are known, then we can determine upper performance bounds; this ideal *oracle* estimation case is useful for algorithm benchmarking purposes [10].

Underdetermined audio source separation is typically addressed by *time-frequency* (TF) *masking*, which assumes that we can transform $x(n)$ by a linear, invertible TF transform so that the sources overlap as little as possible [4]. State-of-the-art methods have the potential to yield sparser representations and superior performance compared to non-adaptive transforms with constant block lengths [9, 10]. Such methods include adaptive, dyadic *lapped orthogonal transforms* (LOTs) [6] and adaptive, non-dyadic LOTs, which give better performance in return for higher computational complexity [7]. We recently introduced *MPEG-like* LOTs, which aim for a trade-off between improving computation time, and decreasing artefacts at window boundaries and improving performance, and evaluated them in oracle contexts [8]. In this paper, we extend this previous work by evaluating them in semi-blind contexts.

2 Time-Frequency Masking

Let us denote by $X(m) = (X_1(m), X_2(m))$ the TF transform of $x(n)$, and let $S(m) = (S_1(m), \ldots, S_J(m))$ be the transform of $s(n)$, where $0 \leq m < N$. We assume that exactly two sources are active at each m because this gives better performance than the simpler *binary masking* case which allows only one active source [1, 10]. The set of both source indices contributing to $X(m)$ is denoted by $\mathcal{J}_m = \{j : S_j(m) \neq 0\}$, and is called the *local activity pattern* at m. Given a particular $\mathcal{J}_m$, Equation (1) then reduces to a determined system:

$$X(m) = A_{\mathcal{J}_m} S_{\mathcal{J}_m}(m) \ , \tag{2}$$

where $A_{\mathcal{J}_m}$ is the 2×2 submatrix of A formed by taking columns A_j, and $S_{\mathcal{J}_m}(m)$ is the subvector of $S(m)$ formed by taking elements $S_j(m)$, whenever $j \in \mathcal{J}_m$. Once $\mathcal{J}_m$ has been estimated for each m we estimate the sources according to the following:

$$\begin{cases} \widehat{S}_j(m) = 0 & \text{if } j \notin \mathcal{J}_m \ , \\ \widehat{S}_{\mathcal{J}_m}(m) = A_{\mathcal{J}_m}^{-1} X(m) & \text{otherwise} \ , \end{cases} \tag{3}$$

where $A_{\mathcal{J}_m}^{-1}$ is the inverse of $A_{\mathcal{J}_m}$ [4]. Finally, we recover the estimated source vector in the time domain $\hat{s}(n)$ by using the inverse transform.

The assumption that exactly two sources are active at each m can be modelled probabilistically by assuming that the source coefficients $S_j(m)$ follow a Laplacian prior distribution, independently and identically for all j and m [1]. In the semi-blind case, the maximum a posteriori solution of (2) is then equivalent to minimising the ℓ^1 norm *cost* of the source coefficients [1] given by the following:

$$C(\widehat{S}) = \sum_{m=0}^{N-1} \sum_{j=1}^{J} |\widehat{S}_j(m)| \ . \tag{4}$$

Then for an orthogonal transform, the estimated semi-blind activity patterns are given by

$$\widehat{\mathcal{J}}_m^{\mathrm{sb}} = \arg\min_{\mathcal{J}_m} \sum_{j=1}^{J} |\widehat{S}_j(m)| \ , \tag{5}$$

which depends implicitly on (3).

3 Adaptive Signal Expansions

Let us now describe how to construct an adapted LOT which better fulfills the sparsity assumption of the sources. This entails forming a *partition* of the domain $\{0, \ldots, N-1\}$ of the mixture channels $x_i(n)$, that is,

$$\lambda = \{(n_k, \eta_k)\} \ , \tag{6}$$

such that

$$0 = n_0 < n_1 < \cdots < n_k < \cdots < n_{K-1} = N-1 \ , \tag{7}$$

where K is the number of partition points. This segments the domain of $x_i(n)$ into adjacent intervals $\mathcal{I}_k = \{n_k, \ldots, n_{k+1} - 1\}$ which should be relatively long over durations which require good frequency resolution, and relatively short over the durations requiring good time resolution. This is achieved by windowing $x_i(n)$ with windows $\beta_k^\lambda(n)$, each of which is supported in $\{n_k - \eta_k, \ldots, n_{k+1} + \eta_{k+1} - 1\}$, thus partly overlapping with its immediately adjacent windows β_{k-1}^λ and β_{k+1}^λ by $2\eta_k$ and $2\eta_{k+1}$ points respectively (see Fig. 1). The *bell parameters*

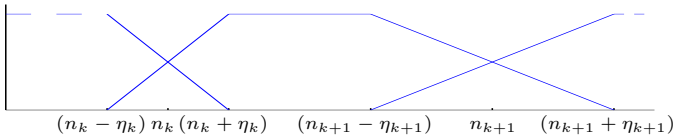

Fig. 1. Schematic representation of window β_k^λ partly overlapping with its adjacent windows β_{k-1}^λ and β_{k+1}^λ.

η_k and η_{k+1} determine how quickly β_k^λ rises and falls on its left and side right sides. To avoid 'double overlapping', these are subject to the constraint

$$n_{k+1} - n_k \geq \eta_{k+1} + \eta_k \ . \tag{8}$$

Note that for $\eta_0 = \eta_{K-1} = 0$ appropriate modifications are needed [6].

For every partition λ we form its associated windows according to

$$\beta_k^\lambda(n) = \begin{cases} r\left(\frac{n-(n_k-\frac{1}{2})}{\eta_k}\right) & \text{if } n_k - \eta_k \le n < n_k + \eta_k \ , \\ 1 & \text{if } n_k + \eta_k \le n < n_{k+1} - \eta_{k+1} \ , \\ r\left(\frac{(n_{k+1}-\frac{1}{2})-n}{\eta_{k+1}}\right) & \text{if } n_{k+1} - \eta_{k+1} \le n < n_{k+1} + \eta_{k+1} \ , \\ 0 & \text{otherwise} \ , \end{cases} \tag{9}$$

where the *bell function* $r(t)$ satisfies $r^2(t) + r^2(-t) = 1$ for $-1 \le t \le 1$, $r(t) = 0$ for $t < -1$ and $r(t) = 1$ for $t > 1$, where t is real-valued, and also satisfies various regularity properties; in practice, we use a sine bell [6].

The *local cosine basis* associated with $\mathcal{I}_k$ is

$$\mathcal{B}_k^\lambda = \left\{ \beta_k^\lambda \sqrt{\frac{2}{n_{k+1} - n_k}} \cos\left[\pi \left(f + \frac{1}{2}\right) \frac{n - (n_k - \frac{1}{2})}{n_{k+1} - n_k}\right] \right\}_{0 \le f < n_{k+1} - n_k} , \tag{10}$$

where the index m in Sect. 2 is now expressed as $m = (k, f)$, where f indexes the 'frequency'. The basis B^λ spanning the space of signals of length N, for the partition λ, is given by $B^\lambda = \bigcup_{k=0}^{K-1} \mathcal{B}_k^\lambda$. Our aim is to find, of all admissible partitions $\lambda \in \Lambda$, the partition which determines the *best orthogonal basis* (BOB) for representing signals of length N. The set of all candidate bases is called the *library* and is given by $\mathcal{L} = \bigcup_{\lambda \in \Lambda} B^\lambda$.

4 Fast and Flexible Partitioning Schemes

For any additive function C, we can use dynamic programming to determine the BOB which minimises $C(\widehat{S})$ over all $B^\lambda \in \mathcal{L}$ [3,6]. Such algorithms jointly estimate the local activity patterns $\mathcal{J}_m$ according to (5) and find the best orthogonal basis which minimises the ℓ^1 norm given by (4) according to

$$\widehat{\lambda} = \arg\min_{\lambda \in \Lambda} C(\widehat{S}) \ . \tag{11}$$

In previous work [7] we described a *flexible segmentation* (FS) partitioning scheme which admits all possible partitions λ with some 'resolution' L, so that if the signal length N is an integral multiple of L, then each partition point can be written as $n_k = cL$ for $c \ge 0$, and where η_k is subject only to the condition (8). The FS library $\mathcal{L}$ is very large due to a combinatorial explosion between the range of allowed interval lengths, interval onsets and bell parameters, so the computation time is typically very high. To decrease this burden of computational complexity, but still wishing to maintain highly flexible partitioning, we subsequently added some constraints to the FS scheme and introduced the following *MPEG-like* partitioning schemes [8]:

Long-Short (LS) We restrict the range of allowable partitions to admit intervals $\mathcal{I}_k$ of only two lengths, that is, a *long interval* of length L_L and a

short interval of length $L_S = L$, where L_L is an integral multiple of L_S, and $2\eta_k \in \{L_L, L_S\}$. Apart from this restriction of interval lengths and bell parameters, there are no additional constraints, and LS is otherwise the same as FS.

Window Shapes (WS) This is equivalent to LS with the additional constraint that if $\mathcal{I}_k$ is long, then at most one of η_k and η_{k+1} is short. In other words, the four different window shapes admitted (compared to five in LS) correspond to a long window ($2\eta_k = 2\eta_{k+1} = L_L$), a short window ($2\eta_k = 2\eta_{k+1} = L_S$), a long-short *transition window* ($2\eta_k = L_L, 2\eta_{k+1} = L_S$), and a short-long ($2\eta_k = L_S, 2\eta_{k+1} = L_L$) transition window in the MPEG-4 framework.

Onset Times (OT) This is equivalent to LS with the additional constraint if any interval $\mathcal{I}_k$ is long, then n_k must satisfy $n_k = cL_L$ for some integer $c = 0, \ldots, \frac{N}{L_L} - 1$.

WS/OT This scheme imposes both the WS and OT constraints simultaneously.

WS/OT/Successive Transitions (WS/OT/ST) This scheme imposes the WS/OT constraints in addition to disallowing adjacent transition windows, i.e., a transition window must be adjacent to a long window and a short window. This implements the MPEG-4 windowing scheme [5], with the exception that here, we have more freedom in choosing the bell function $r(t)$.

Clearly, the sizes of the libraries become smaller as we impose more constraints.

5 Experiments and Results

We performed two sets of experiments to test our algorithms. Performance is measured through the *signal to distortion ratio* (SDR) [10],

$$\text{SDR [dB]} = 10 \log_{10} \frac{\sum_{n=0}^{N-1} \sum_{j=1}^{J} \left(s_j(n)\right)^2}{\sum_{n=0}^{N-1} \sum_{j=1}^{J} \left(\hat{s}_j(n) - s_j(n)\right)^2} \ . \tag{12}$$

In the first set of experiments, we applied our methods to twenty mixtures in total (ten music mixtures, ten speech mixture), where each mixture each had $J = 3$ sources at a sampling rate of 22.05 kHz, with a resolution of 16 bits per sample, and of length $N = 2^{18}$ (approximately 11.9 s). The sources were mixed according to following mixing matrix:

$$A = \begin{pmatrix} 0.21 & 0.95 & 0.64 \\ 0.98 & 0.32 & 0.77 \end{pmatrix} \ . \tag{13}$$

For each mixture, we performed semi-blind estimations of $s(n)$ for each of the LS, WS, OT, WS/OT and WS/OT/ST partitioning schemes, with long intervals $L_L = 2^c$, where $c \in \{8, \ldots, 11\}$ (12 ms to 93 ms), and short intervals $L_S = 2^c$, where $n \in \{4, \ldots, 9\}$ (0.73 ms to 23 ms). We exclude all long-short combinations with $L_L \leq L_S$. Results are presented in Table 1 where each entry is the average over the twenty different mixtures corresponding to a particular transform scheme with given block lengths. We also compare the MPEG-like schemes to

Table 1. Average results for MPEG-like transforms for semi-blind separation on music and speech mixtures (see text). The baseline (fixed basis, FB) transform scheme yields maximum average SDR of 12.06 dB at $L_L = L_S = 2^{10}$.

Scheme	L_L	L_S					
		2^4	2^5	2^6	2^7	2^8	2^9
LS	2^8	10.45	10.50	10.51	10.55	-	-
	2^9	11.72	11.71	11.72	11.72	11.79	-
	2^{10}	12.14	12.10	12.19	12.16	12.23	12.29
	2^{11}	11.70	11.59	11.73	11.77	11.92	**12.34**
WS	2^8	10.45	10.51	10.52	10.55	-	-
	2^9	11.76	11.71	11.74	11.74	11.80	-
	2^{10}	12.16	12.14	12.18	12.16	12.23	**12.28**
	2^{11}	11.62	11.66	11.69	11.75	11.91	12.22
OT	2^8	10.68	10.66	10.65	10.64	-	-
	2^9	11.83	11.83	11.85	11.85	11.83	-
	2^{10}	12.07	12.07	12.07	12.06	12.15	12.19
	2^{11}	11.65	11.56	11.60	11.61	11.86	**12.29**
WS/OT	2^8	10.68	10.67	10.66	10.64	-	-
	2^9	11.84	11.83	11.85	11.85	11.83	-
	2^{10}	12.07	12.07	12.08	12.08	12.16	12.20
	2^{11}	11.62	11.56	11.59	11.61	11.83	**12.29**
WS/OT/ST	2^8	10.69	10.68	10.67	10.64	-	-
	2^9	11.84	11.84	11.85	11.85	11.85	-
	2^{10}	12.05	12.04	12.06	12.08	12.16	12.21
	2^{11}	11.57	11.52	11.53	11.55	11.77	**12.28**

the baseline *fixed basis* (FB) transform (where $L_L = L_S$ and $2\eta_k = L_L$ for all k) and find that the maximum average SDR is 12.06 dB at $L_L = L_S = 2^{10}$.

For the second set of experiments, we indicate the performance achievable on particular types of mixtures. We applied the best transform scheme as determined by Table 1 (LS) to each instantaneous mixture in the *dev1* data set of the *Signal Separation Evaluation Campaign* (SiSEC 2008)[3]. These optimal semi-blind results are presented in Table 2; also shown are oracle estimation results, where the L_L and L_S which give best results were determined in previous work [8]. Oracle results are computed by jointly determining the local activity patterns $\mathcal{J}_m$ and the best orthogonal basis $B^\lambda \in \mathcal{L}$ which maximise the SDR given by (12), given knowledge of the reference sources [9].

Table 2. Results for LS scheme for semi-blind and oracle separation on SiSEC 2008 data (see text).

Mixture	J	Semi-blind			Oracle		
		L_L	L_S	Av. SDR [dB]	L_L	L_S	Av. SDR [dB]
3 Female Speakers	3	2^9	2^5	10.35	2^{10}	2^4	24.09
4 Female Speakers	4	2^{11}	2^9	7.04	2^{10}	2^4	18.61
3 Male Speakers	3	2^9	2^9	8.41	2^{10}	2^4	18.56
4 Male Speakers	4	2^{10}	2^9	5.62	2^{10}	2^4	14.37
Music with No Drums	3	2^{10}	2^7	16.33	2^{10}	2^4	34.26
Music with Drums	3	2^9	2^4	11.95	2^{10}	2^4	28.06

6 Discussion

For the results in Table 1, the best average SDR is approximately 12.3 dB for each transform scheme. Previous results demonstrated oracle performance of 23–25 dB, but the differences between the two cases are not surprising; the oracle estimation criterion is the same as the performance measurement criterion (SDR), whereas the semi-blind estimation criterion (ℓ^1 norm) is different.

The greatest variability in average SDR occurs with changing the long interval length L_L. The SDR improvements in the demonstrated range of 1–2 dB may be significant in high fidelity applications. Varying L_S or changing transform scheme has a much smaller effect on performance, in contrast to previous oracle results, where performance naturally decreases as the partitioning schemes get more restrictive and their respective libraries becoming smaller.

[3] Available online at `http://sisec.wiki.irisa.fr/tiki-index.php`

In each case in Table 1, the best average SDR is achieved at the *greatest* length for the short intervals ($L_S = 2^9$). In contrast, Table 2 shows individual, rather than average, results. Previous oracle results for the LS and WS schemes show that the best average SDR was obtained at the *least* length for the short intervals ($L_S = 2^4$), where we suggested that a library which allows fine-grained placement of the long windows improves performance [8]. The current ℓ_1 criterion does not achieve this, but a semi-blind criterion which admits such fine-grained placement will be a good step towards closing the performance gap between semi-blind and oracle performance. This claim is strengthened by noting that the average SDR improvement yielded by adaptive schemes compared to FB is in the order of 0.3 dB in the semi-blind case, and 1–2 dB in oracle contexts.

7 Conclusions and Further Work

We demonstrated average SDR performance of 12–13 dB on mixtures of music and speech signals by extending our adaptive signal decomposition schemes to the semi-blind case. Table 1 suggests that optimal results are obtained when both L_L and L_S are long, but this requires further investigation. Further work includes extending this technique to the fully blind case. Preliminary experiments on mixing matrix estimation with the SiSEC 2008 data sets using histogram-based methods [2] have shown very promising results, and we intend to incorporate that framework into our adaptive transform schemes.

References

1. Bofill, P., Zibulevsky, M.: Underdetermined blind source separation using sparse representations. Signal Process. 81(11), pp. 2353–2362 (2001)
2. Bofill, P.: Identifying single source data for mixing matrix estimation in instantaneous blind source separation. In: Proc. ICANN 2008, pp. 759–767 (2008)
3. Huang, Y., Pollak, I., Bouman, C. A., Do, M. N.: Best basis search in lapped dictionaries. IEEE Trans. Signal Process. 54(2), pp. 651–664 (2006)
4. Gribonval, R.: Piecewise linear source separation. In: Proc. SPIE (Wavelets X), 5207, pp. 297–310 (2003)
5. ISO: Information technology—Coding of audio-visual objects—Part 3: Audio (ISO/IEC 14496-3:2005). ISO, Geneva, Switzerland (2005)
6. Mallat, S.: A Wavelet Tour of Signal Processing. Second ed., Academic Press (1999)
7. Nesbit, A., Plumbley, M. D., Vincent, E.: Oracle evaluation of flexible adaptive transforms for underdetermined audio source separation. In: Proc. ICArn 2008, pp. 17–20 (2008)
8. Nesbit, A., Vincent, E., Plumbley, M. D.: Benchmarking flexible adaptive time-frequency transforms for underdetermined audio source separation. Submitted to: ICASSP 2009 (2009)
9. Vincent, E., Gribonval, R.: Blind criterion and oracle bound for instantaneous audio source separation using adaptive time-frequency representations. In: Proc. WASPAA2007, pp. 110–113 (2007)
10. Vincent, E., Gribonval, R., Plumbley, M. D.: Oracle estimators for the benchmarking of source separation algorithms. Signal Process. 87(8), pp. 1933–1950 (2007)