Oracle evaluation of flexible adaptive transforms for underdetermined audio source separation

Andrew Nesbit, Mark Plumbley, Emmanuel Vincent

the time domain index, n, range as $0 \leq n < N$. We concentrate on such a model because it can provide a useful approximation to stereo audio signals consisting of $J > 2$ sources (underdetermined), mixed using a *panned mono* mixing technique (instantaneous). If A is unknown, then the problem is called *blind*, otherwise we call it *semi-blind*.

1.1 Time-Frequency Masking

In the underdetermined case, even if we know or have estimated A, matrix inversion techniques will not give a unique solution for $s(n)$. However, we can use techniques based on time-frequency (TF) masking, which transform the observed $x(n)$ using linear, invertible, TF transforms [12]. This preserves the mixing structure of (1) to give

$$[X_i(k,f)]_{1 \leq i \leq 2} = A[S_i(k,f)]_{1 \leq j \leq J} \qquad (2)$$

where $(k,f) \in \Gamma$ is a time-frequency index pair, the components of which index the block and frequency, respectively.

Denote by $J'_{k,f}$ the assumed number of active (non-zero) source coefficients at the TF index (k,f). Then $\mathcal{J}_{k,f} = \{j : S_j(k,f) \neq 0\}$ is the set of all $J'_{k,f}$ sources contributing to $[X_i(k,f)]_{1 \leq i \leq 2}$, and is called the *local activity pattern* at (k,f). Then (2) reduces to

$$[X_i(k,f)]_{1 \leq i \leq 2} = A_{\mathcal{J}_{k,f}}[S_i(k,f)]_{j \in \mathcal{J}_{k,f}} \qquad (3)$$

where $A_{\mathcal{J}_{k,f}}$ is the $2 \times J'_{k,f}$ submatrix of A formed by taking columns A_j, and $[S_i(k,f)]_{j \in \mathcal{J}_{k,f}}$ is formed by taking rows of $[S_j(k,f)]_{1 \leq j \leq J}$, whenever $j \in \mathcal{J}_{k,f}$. A fundamental assumption of the TF representation is that $J'_{k,f} \leq 2$, in other words, that it admits a sparse representation. Then, in the ideal case, (3) can be solved for each (k,f) independently according to

$$\begin{cases} S_j(k,f) = 0 & \text{if } j \notin \mathcal{J}_{k,f} \\ [S_j(k,f)]_{j \in \mathcal{J}_{k,f}} = A^+_{\mathcal{J}_{k,f}}[X_i(k,f)]_{1 \leq i \leq 2} & \text{otherwise} \end{cases}$$
$$\qquad (4)$$

where $\mathcal{J}_{k,f}$ is an estimate of $\mathcal{J}_{k,f}$ and $A^+_{\mathcal{J}_{k,f}}$ denotes the (Moore-Penrose) pseudoinverse of $A_{\mathcal{J}_{k,f}}$ [3]. Time frequency masking can then be interpreted as the problem of estimating local activity patterns. While binary masking

Andrew Nesbit and Mark D. Plumbley
Department of Electronic Engineering
Queen Mary, University of London
Mile End Road, London, E1 4NS
United Kingdom
andrew.nesbit@elec.qmul.ac.uk
mark.plumbley@elec.qmul.ac.uk

Emmanuel Vincent
METISS Group
IRISA-INRIA
Campus de Beaulieu, 35042 Rennes Cedex
France
emmanuel.vincent@irisa.fr

ORACLE EVALUATION OF FLEXIBLE ADAPTIVE TRANSFORMS FOR UNDERDETERMINED AUDIO SOURCE SEPARATION

ABSTRACT

We describe and apply a flexible, adaptive cosine packet transform to separate audio sources from instantaneous, underdetermined audio mixtures by time-frequency masking. Previously studied adaptive transform schemes have two main drawbacks: the signal can only be partitioned into dyadic intervals, and the profiles of the overlapping windows are often very short, thus tapering off very quickly. The novel aspects of our new approach are that it admits a much larger library of admissible orthogonal bases, and thus does not require dyadic segmentation and alleviates border artifacts at window boundaries.

Oracle estimation, which determines experimental upper performance bounds of our techniques, demonstrates potential performance improvements of up to 3.0 dB SDR, when compared with fixed-basis transforms such as the short-time Fourier transform and modified discrete cosine transform, and the previously studied adaptive cosine packet decomposition scheme.

Keywords: source separation, oracle estimators, adaptive transforms.

1 INTRODUCTION

The aim of audio source separation is to estimate a set of simultaneously active audio *sources* from a set of observed *mixtures* of those sources [9]. Let us consider the case of the two-channel, *underdetermined*, *instantaneous* mixture, and form the following time domain model of the mixing process:

$$x(n) = As(n), \qquad (1)$$

where $x(n) = [x_i(n)]_{1 \leq i \leq 2}$ is a column vector representing the two-channel mixture signal, $s(n) = [s_i(n)]_{1 \leq j \leq J}$ is a column vector of J source signals, and $A = [a_{i,j}]_{1 \leq i \leq 2; 1 \leq j \leq J}$ is the matrix of *mixing parameters*. Let

($J'_{k,f} = 1$) is now well understood [12], efficient estimation of activity patterns with $J'_{k,f} \leq J$ case is currently an open question.

1.2 The problem

Commonly used transforms which can be used to estimate $[S(k,f)]_{1 \leq j \leq J}$ according to (4) include the short-time Fourier transform (STFT) [12], the modified discrete cosine transform (MDCT) [2] and adaptive cosine packets [7].

It has been shown that signal decompositions using adaptive cosine packets have the potential to give superior performance [6, 10] to the STFT and MDCT. However, problems with these approaches include the necessity of windowing the signal over dyadic segments, and the presence of border effects at window boundaries. In Section 2, we describe a scheme which alleviates these problems, and in Section 4, we show that the new scheme has the potential to outperform the old one.

2 FLEXIBLE, ADAPTIVE SIGNAL REPRESENTATIONS

One of the motivations for the use of orthogonal, adaptive transforms is that they have the potential to represent the sources more sparsely than fixed-basis or overcomplete transforms (such as the STFT). Previous studies have examined the benefits afforded by adaptive cosine packet transforms in (semi-)blind [7, 8] and oracle contexts [6, 8].

The aim is to partition the signal using overlapping windows of variable length. This defines an orthonormal transform adapted to the time-varying characteristics of the signal. Ideally we obtain longer windows over intervals requiring fine frequency resolution, at the expense of coarser time resolution, and shorter windows over intervals with broadband frequency content, giving finer time resolution.

2.1 Mathematical Definition of the Bases

The following exposition follows the style and notation developed in previous work on adaptive cosine packets [4, 5]. Let λ denote a partition of the entire signal:

$$\lambda = \{(n_k, \eta_k)\}_{0 \leq k \leq K_\lambda}$$

where

$$n_0 = 0 < n_1 < \cdots < n_{K_\lambda - 1} < n_{K_\lambda} = N$$

and where, for each n_k, there is an associated bell parameter, η_k, which is the half-length of the overlap interval between successive windows, and which defines the shape of the overlapping windows such that we have

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

The important thing about this development is that the bell parameters, η_k, are not necessarily all equal. In contrast to this, previous work using adaptive cosine packet transforms imposed the constraint that $\eta_k = \eta$ is constant

across all $k = 1, \ldots, K_\lambda - 1$. For the $k = 0, K_\lambda - 1$ cases, appropriate border modifications need to be made [5].

For every k and $k + 1$ associated with the partition, λ, we form an interval $[n_k - \eta_k, n_{k+1} + \eta_{k+1}]$, to which the restriction of the signal is made through the use of a window function:

$$\beta_{n_k, n_k+1}^{\eta_k, \eta_{k+1}}(n) =$$

$$\begin{cases} r\left(\frac{n - (n_k - 1/2)}{\eta_k}\right) & \text{if } n_k - \eta_k \leq n < n_k + \eta_k \\ 1 & \text{if } n_k + \eta_k \leq n < n_{k+1} - \eta_{k+1} \\ r\left(\frac{(n_{k+1} - 1/2) - n}{\eta_{k+1}}\right) & \text{if } n_{k+1} - \eta_{k+1} \leq n < n_{k+1} + \eta_{k+1} \\ 0 & \text{otherwise.} \end{cases}$$

The bell function r satisfies $r^2(t) + r^2(-t) = 1$ for $-1 \leq t \leq 1$, $r(t) = 0$ for $t < -1$, and $r(t) = 1$ for $t > 1$, where t is real-valued, and also satisfies various regularity properties [5]. The bell parameters η_k and η_{k+1} determine how quickly the window monotonically rises on its left side and monotonically falls on its right side. Although there are many windows which satisfy these constraints, in practice, we use a sine window [5]. The local cosine basis spanning the signal space for this interval is then given by

$$\mathcal{B}_{n_k, n_k+1}^{\eta_k, \eta_{k+1}} = \left\{ \beta_{n_k, n_k+1}^{\eta_k, \eta_{k+1}}(n) \sqrt{\frac{2}{n_{k+1} - n_k}} \right.$$

$$\left. \times \cos\left[\pi \left(f + \frac{1}{2} \right) \frac{n - (n_k - 1/2)}{n_{k+1} - n_k} \right] \right\}_{0 \leq f < n_{k+1} - n_k}$$

where f is the discrete frequency index.

Now we are finally in a position to construct the orthonormal basis, B^λ, associated with this particular λ, for the space of signals of length N:

$$B^\lambda = \bigcup_{k=0}^{K_\lambda - 1} \mathcal{B}_{n_k, n_{k+1}}^{\eta_k, \eta_{k+1}}.$$

This basis is only one of many possibilities. Since our aim is to find the best basis, we will consider *all* admissible partitions, $\lambda \in \Lambda$, each of which determines a different orthonormal basis. Thus we obtain a *library* of possible cosine packet bases for this space of signals of length N:

$$\mathcal{L} = \bigcup_{\lambda \in \Lambda} B^\lambda.$$

2.2 Computing the Best Basis

Our aim is to find that $B^\lambda \in \mathcal{L}$ which gives the best representation of our signal by minimising an additive cost function whose value is inversely related to separation performance. (In Section 3 we will define a cost function for oracle estimation.) Denote by $C_{n_k, n_k+1}^{\eta_k, \eta_{k+1}}$ the cost of representing the signal in the interval $[n_k - \eta_k, n_{k+1} + \eta_{k+1}]$ over the basis $\mathcal{B}_{n_k, n_k+1}^{\eta_k, \eta_{k+1}}$. If the cost function is additive, then the overall cost of representing the signal over the basis $B^\lambda = \bigcup_{k=0}^{K_\lambda - 1} \mathcal{B}_{n_k, n_{k+1}}^{\eta_k, \eta_{k+1}}$ is given by

$$C_{0,N}^{0,0} = \sum_{k=0}^{K_\lambda - 1} C_{n_k, n_{k+1}}^{\eta_k, \eta_{k+1}}.$$

Past research on the use of adaptive cosine packet transforms involves restricting the set of admissible segmentations so that, in particular, the length of each interval $[n_k, n_{k+1}]$ as well as its end points are powers of two for all k, and so that all bell parameters $\eta_k = \eta = 2L$ for $k = 1, \ldots, K_\lambda - 1$ [6, 7, 8]. This has the desirable effect that we can use the computationally efficient Coifman-Wickerhauser (CW) algorithm [1] to determine the best orthogonal basis with minimum cost $C_{0,N}^{0,0}$. On the other hand, this also severely restricts the range of admissible partitions, and hence the library from which we choose the best basis is much smaller. It also causes distortions in the estimated sources due to windowing artifacts because if η is small, then very short overlaps will occur even between two relatively long adjacent windows [7]. To overcome these problems, we employ an alternative, flexible segmentation (FS) algorithm, also based on dynamic programming [4, 11]. Whereas for the CW algorithm to be applicable, the library must be representable as a complete dyadic tree, the FS algorithm is much more lenient. It permits time segmentations of resolution L, so that a signal of length N is a multiple of L, and each partition point can be written as $n_k = cL$ for some integer $c \geq 0$. Furthermore, provided that both L and N are powers of two, the FS library is a superset of the CW library.

3   ORACLE ESTIMATORS FOR AUDIO SOURCE SEPARATION

Oracle estimation allows us to judge the difficulty of separating the sources from a given mixture and to gain insight into the upper performance bounds of our class of separation algorithms. As it depends on knowing the reference source signals, s, and the mixing matrix, A, it is intended to be used for *algorithm evaluation* rather than for practical (semi-)blind separation applications. The aim is to determine those $J'_{k,f}$ and $\mathcal{J}_{k,f}$ which give the best possible separation performance by optimising against some performance criterion [10].

The oracle estimate of $s(n)$ is the $\hat{s}(n)$ which minimises a distortion measure such as

$$C_{0,N}^{0,0} = \sum_{n=0}^{N-1} \sum_{j=1}^{J} \left(\hat{s}_j(n) - s_j(n) \right)^2, \qquad (5)$$

such that $\hat{s}(n)$ has been estimated by applying (4) in a particular basis B^λ. The advantages of defining C in this way are that minimising it is equivalent to maximising the signal to distortion ratio (SDR [dB]), given by

$$\text{SDR} = 10 \log_{10} \frac{\sum_{n=0}^{N-1} \sum_{j=1}^{J} \left(s_j(n) \right)^2}{C_{0,N}^{0,0}},$$

which we will use to evaluate our methods; and that it satisfies the additivity constraints required for computing the best orthogonal basis (Section 2.2).

For signals represented by an orthonormal transform,

(5) is equal to the following [8]:

$$C_{0,N}^{0,0} = \sum_{k=0}^{K_\lambda - 1} C_{n_k, n_k+1}^{\eta_k, \eta_{k+1}}$$
$$= \sum_{k=0}^{K_\lambda - 1} \sum_{f=0}^{n_{k+1} - n_k - 1} \sum_{j=1}^{J} \left(\hat{S}_j(k, f) - S_j(k, f) \right)^2,$$

where the transform coefficients are computed in the basis B^λ. It is clear that minimising $C_{0,N}^{0,0}$ is equivalent to minimising at each (k, f) independently, by computing oracle local activity patterns:

$$\mathcal{J}_{k,f}^{\text{ora}} = \arg\min_{\mathcal{J}_{k,f} \in \mathcal{P}_{k,f}} \sum_{j=1}^{J} \left(\hat{S}_j(k, f) - S_j(k, f) \right)^2$$

where $\hat{S}_j(k, f)$ on the right hand side is given by (4), and $\mathcal{P}_{k,f}$ is the set of all possible activity patterns subject to $J'_{k,f}$. If $J'_{k,f}$ is small then an exhaustive search over all $\mathcal{J}_{k,f} \in \mathcal{P}_{k,f}$ is computationally feasible.

The STFT is often used in time-frequency masking, but because it is non-orthogonal, it is computationally infeasible to determine the optimal oracle activity patterns. We are therefore restricted to computing *near-optimal* oracle activity patterns, as in [10].

4   EXPERIMENTS AND RESULTS

We test and compare the various methods and transforms on a mixture of $J = 4$ four musical sources. As we had access to the original multitracked data, we were able to synthesise instantaneous mixtures, with $I = 2$, to simulate a *panned mono* mixing process, using the following mixing matrix:

$$A = \left(\begin{array}{cccc} \cos(\frac{\pi}{16}) & \cos(\frac{3\pi}{16}) & \cos(\frac{5\pi}{16}) & \cos(\frac{7\pi}{16}) \\ \sin(\frac{\pi}{16}) & \sin(\frac{3\pi}{16}) & \sin(\frac{5\pi}{16}) & \sin(\frac{7\pi}{16}) \end{array} \right)$$

The pitched sources were harmonically related so that overlapping harmonics between different sources were expected. To ease computation time, we downsampled from 44.1 kHz to 22.05 kHz, kept at a resolution of 16 bits per sample. The extract was of length 2^{18} samples (approximately 11.9 s).

We allow the number of active sources at each time-frequency index to range as $J'_{k,f} \leq 2$. Previous experiments using oracle estimators on adaptive cosine packet transforms (using the CW algorithm) have shown that this gives significantly higher performance than the $J'_{k,f} = 1$ and $J'_{k,f} = 2$ cases, and hence, a better indication of what time-frequency masking is potentially capable of achieving [6].

The resolution for the adaptive cosine packet transforms determined by the CW and FS algorithms was set to $L = 2^8$. We tested these transform methods with a fixed bell parameter, $\eta = 2^8$, and, for the FS method, also tested a range of bell parameters, $\eta = c \cdot 2^8$, where $c = 1, \ldots, 16$, to take advantage of the large library it offers. For the STFT and MDCT fixed-basis transforms, the window overlap and block length parameters which gave

Trans.	η	L	Av. SDR [dB]
STFT	2^{12}	2^{13}	13.8
MDCT	2^{10}	2^{11}	14.6
CW	2^8	2^8	15.9
FS	2^8	2^8	16.3
	$c \cdot 2^8 \ (c = 1, \ldots, 16)$	2^8	16.8

Table 1: Results of oracle estimation. For the CW and FS transforms, the parameter η indicates the range of bell parameters used in adapting to the signal, and L is the resolution parameter. For the STFT and MDCT transforms, the meanings of η and L are slightly different, and indicate window overlap and fixed block size. The average SDR of all extracted sources is given in dB.

the best average SDR in previous work were chosen a priori [6].

Results are presented in Table 1. Informal experimentation indicates that the relative performance improvements offered by the FS algorithm over the CW algorithm, and over the MDCT and STFT transforms, are typical of other mixtures of music sources as well.

5 DISCUSSION

Although we have presented results for only one mixture, the relative performance differences between the transforms are fairly typical for instantaneous, two-channel mixtures of four harmonically related instruments.

Although the performance improvements are modest, ranging from between several tenths to a whole decibel, the increase in computation time far outweighed any performance improvement as judged by the SDR. Decomposing the mixtures using the CW algorithm with a fully flexible windowing scheme took in the order of days to execute, whereas the same algorithm with only one possible bell parameter generally takes in the order of minutes. In our implementation, approximately three quarters of the computation time is spent decomposing the segments in local cosine bases, with the remaining quarter spent determining the oracle masks.

6 CONCLUSIONS AND FUTURE WORK

Using the FS algorithm to find the best orthogonal basis for oracle estimation of audio sources leads to a modest improvement of about 1 dB over more well known methods. However, for practical scenarios this performance increase is outweighed by the large increase in computation required to determine the best partition of the signal. Future work include investigating the use of algorithms such as beam search to find a suboptimal orthogonal basis using a flexible segmentation scheme, and listening tests to determine subjective, perceptual differences between the different transform schemes.

References

[1] Ronald R. Coifman and Mladen Victor Wickerhauser. Entropy-based algorithms for best basis selection. *IEEE Transactions on Information Theory*, 38(2):713–718, March 1992.

[2] M. Davies and N. Mitianoudis. Simple mixture model for sparse overcomplete ICA. *IEE Proceedings on Vision, Image and Signal Processing*, 151 (1):35–43, February 2004.

[3] Rémi Gribonval. Piecewise linear source separation. In Michael A. Unser, Akram Aldroubi, and Andrew F. Laine, editors, *Proceedings of the SPIE (Wavelets: Applications in Signal and Image Processing X)*, volume 5207, pages 297–310. SPIE, WA, USA, November 2003.

[4] Yan Huang, Ilya Pollak, Charles A. Bouman, and Minh N. Do. Best basis search in lapped dictionaries. *IEEE Transactions on Signal Processing*, 54(2): 651–664, February 2006.

[5] Stéphane Mallat. *A Wavelet Tour of Signal Processing*. Academic Press, second edition, 1999.

[6] Andrew Nesbit and Mark D. Plumbley. Oracle estimation of adaptive cosine packet transforms for underdetermined audio source separation. In *Proceedings of the 2008 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2008)*, Las Vegas, NV, USA, March 2008.

[7] Andrew Nesbit, Mark D. Plumbley, and Mike E. Davies. Audio source separation with a signal-adaptive local cosine transform. *Signal Processing*, 87(8):1848–1858, August 2007.

[8] Emmanuel Vincent and Rémi Gribonval. Blind criterion and oracle bound for instantaneous audio source separation using adaptive time-frequency representations. In *Proceedings of the 2007 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA2007)*, New Paltz, NY, USA, October 2007.

[9] Emmanuel Vincent, Maria G. Jafari, Samer A. Abdallah, Mark D. Plumbley, and Mike E. Davies. Blind audio source separation. Technical Report C4DM-TR-05-01, Department of Electronic Engineering, Queen Mary, University of London, November 2005.

[10] Emmanuel Vincent, Rémi Gribonval, and Mark D. Plumbley. Oracle estimators for the benchmarking of source separation algorithms. *Signal Processing*, 87(8):1933–1950, August 2007.

[11] Zixiang Xiong, Kannan Ramchandran, Cormac Herley, and Michael T. Orchard. Flexible tree-structured signal expansions using time-varying wavelet packets. *IEEE Transactions on Signal Processing*, 45(2): 333–345, February 1997.

[12] Özgür Yılmaz and Scott Rickard. Blind separation of speech mixtures via time-frequency masking. *IEEE Transactions on Signal Processing*, 52 (7):1830–1847, July 2004.