

# A SIGNAL-ADAPTIVE LOCAL COSINE TRANSFORM FOR SOURCE SEPARATION BY TIME-FREQUENCY MASKING

Andrew Nesbit    Mark D. Plumbley  
Department of Electronic Engineering  
Queen Mary, University of London  
{andrew.nesbit, mark.plumbley}  
@elec.qmul.ac.uk

Mike E. Davies  
IDCOM & Joint Research Institute  
for Signal and Image Processing  
University of Edinburgh  
mike.davies@ed.ac.uk

## ABSTRACT

Time-frequency masking is often used for source separation of underdetermined audio mixtures. It depends on the fact that the sources can be represented disjointly in some transform domain. The focus of this paper is on demixing sources from instantaneous, two-channel mixtures by binary masking. We investigate trees of local cosine bases from which a suitable transform may be generated—the best basis is chosen by a computationally efficient algorithm and is adaptively selected to match the time-varying characteristics of the signal. Our heuristically motivated cost function maximises the energy of the transform coefficients associated with each estimated source.

Finally, we evaluate our proposed transform by comparing it against two well-known transforms: the short-time Fourier transform and the modified discrete cosine transform. We assume that the mixing parameters are known. Our results show that in some cases, our method can give better results than these fixed-basis representations.

**Keywords:** Audio source separation, local cosine bases

## 1 INTRODUCTION

This work deals with audio source separation of *underdetermined, instantaneous, two-channel* mixtures of  $n > 2$  audio sources:

$$x_i = \sum_{j=1}^n a_{i,j} s_j \quad (1)$$

where  $s_j$  is the  $j$ th source,  $x_i$  is the  $i$ th mixture,  $a_{i,j}$  is the positive real amplitude (mixing parameter) of the  $j$ th source in the  $i$ th mixture (observation), and  $1 \leq j \leq n$  and  $i = 1, 2$ . If the mixing parameters  $a_{i,j}$  are unknown, the problem is called *blind*.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page.

©2006 The University of Liverpool

The blind source separation problem may be split conceptually into two successive subproblems [12]. Estimation of the  $a_{i,j}$  constitutes the *identification* phase, while extraction of the  $s_j$ , to yield estimated sources  $\hat{s}_j$  is called the *filtering* phase. This paper concentrates on the demixing (filtering) phase. We assume that the mixing parameters are known, or have already been estimated, for example, by forming a histogram of mixing parameter estimates [13] or by clustering in high-dimensional spaces [4].

The structure of this paper is as follows: In Section 2 we introduce the method of source separation by time-frequency masking. Section 3 introduces the cosine packet (CP) tree approach, together with our proposed cost functions for its use in source separation. In Section 4 the proposed method is evaluated and compared to the short-time Fourier transform (STFT) and the modified discrete cosine transform (MDCT), and is followed by a discussion of further work and conclusions.

## 2 TIME-FREQUENCY MASKING

One requirement of time-frequency masking methods is that the sources have a disjoint representation [13]. We apply a real- or complex-valued linear transform  $T$  to the mixtures  $x_1$  and  $x_2$  to give a pair of transformed mixtures  $\tilde{x}_1 = Tx_1$  and  $\tilde{x}_2 = Tx_2$ . This represents the transformed sources with the following same mixing structure as in Equation (1):

$$\tilde{x}_i = \sum_{j=1}^n a_{i,j} \tilde{s}_j \quad (2)$$

where  $\tilde{x}_i$  and  $\tilde{s}_i$  are the respective transforms of the signals  $x_i$  and  $s_i$ .

If  $T$  transforms the mixtures so that the sources have (approximately) disjoint representations, then at any given point in the time-frequency plane, indexed by  $\gamma$ , the contribution from at most one source,  $s_j$  will dominate. Equation (2) then reduces to the following simple linear system:

$$\tilde{x}_i(\gamma) = a_{i,j} \tilde{s}_j(\gamma). \quad (3)$$

at all time-frequency indices  $\gamma$  where  $s_j$  dominates. This means that if we know  $a_{i,j}$  and can estimate the points at which  $s_j$  dominates, then we can extract the source by forming a binary time-frequency mask (Section 2.1).

To obtain a disjoint representation we use *sparse* transforms [9, 10], in which most coefficients are very close to zero and only a few coefficients are large. The probability that two or more sources are simultaneously large is very small. This means that a sparse transform will represent the mixtures with approximately disjoint support in the transform domain. An initial study of the performance effects of different transforms appears in [10].

Commonly used sparse transforms for time-frequency masking include the STFT [13] and the MDCT [3, 10]. Each of these transforms has a fixed basis set. In order to better match the time-varying characteristics of the mixtures and sources, we propose an adaptive transform [7], whose bases and window lengths are adapted to the input signals (Section 3.1). Examples of this sort of approach include ICA-like techniques which use local cosine bases and wavelet packets to represent the signals [5], and clustering of wavelet packet coefficients [6].

## 2.1 Filtering the Sources

Given that we know or have estimated the ratios of mixing coefficients, individual sources can be estimated from  $\tilde{x}_1$  and  $\tilde{x}_2$  by constructing *binary time-frequency masks*. The ratio of mixing coefficients for the  $j$ th source can be associated with an angle

$$\theta_j = \arctan\left(\frac{a_{2j}}{a_{1j}}\right) \quad (4)$$

where the inverse tangent is computed in the first quadrant of the plane. If the transform  $Ts_j = \tilde{s}_j$  of the  $j$ th source is sparse and real-valued, then the coefficients tend to cluster along the line defined by  $\theta_j$  in  $\tilde{x}_1$ - $\tilde{x}_2$  space. A binary time-frequency mask  $M_{\theta_j, u}$  captures the coefficients which fall ‘close’ to this line, that is, all coefficients whose time-frequency points dominate at that angle, and discards all others. For simplicity we define our masks as

$$M_{\theta_j, u} = \begin{cases} 1 & \text{if } \theta_j - \frac{u}{2} < \arctan\left(\frac{\tilde{x}_2}{\tilde{x}_1}\right) < \theta_j + \frac{u}{2} \\ 0 & \text{otherwise.} \end{cases} \quad (5)$$

This mask  $M_{\theta_j, u}$  estimates the mixture coefficients carrying most of the energy for source  $j$ , for the given manually set symmetric threshold  $u$ .

Once the masks have been constructed, they can be used to determine  $\hat{\tilde{s}}_j$ , an estimation of the transformed  $j$ th source:

$$\hat{\tilde{s}}_j = M_{\theta_j, u} \cdot (\tilde{x}_1 \cos \theta_j + \tilde{x}_2 \sin \theta_j). \quad (6)$$

Finally, we apply the inverse transform to recover the time-domain sources  $s_j$ .

## 3 ADAPTING THE REPRESENTATION

Local cosine transforms, in which the transform adapts to the input mixtures, can give 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). This may give a sparser representation than one obtained by a fixed-basis transform such as the STFT or MDCT.

### 3.1 Trees of local cosine bases

The basis functions of the cosine packet transform [7] are defined over dyadic-length (powers of 2) intervals  $[c_{p,d}, c_{p+1,d}]$ . The endpoints are given by

$$c_{p,d} = 2^{-d}Np - \frac{1}{2} \quad (7)$$

where  $N$  is the length of the (time-domain) signal, and these define a binary tree structure where the depth of a node is given by  $d$  up to a maximum depth  $D$  ( $0 \leq d \leq D$ ), and the position of a node at level  $d$  is given by  $p$  ( $0 \leq p < 2^d$ ). A pair of indices  $(p, d)$  corresponds to a node in the tree, and identifies a signal space spanned by an orthogonal *local cosine basis*:

$$\left\{ w_{p,d}[n] \sqrt{\frac{2}{2^{-d}N}} \cos \left[ \pi \left( k + \frac{1}{2} \right) \frac{n - c_{p,d}}{2^{-d}N} \right] \right\} \quad (8)$$

where  $0 \leq k < 2^{-d}N$  indexes the functions in the basis. The smooth window  $w_{p,d}$  localises the basis functions over the interval  $[c_{p,d}, c_{p+1,d}]$ , partly overlaps with its immediately adjacent windows  $w_{p-1,d}$  and  $w_{p+1,d}$ , and satisfies other special properties [7]. We use this adaptive framework to our advantage to attempt represent the sources more disjointly. Moreover, this tree structure offers a computationally efficient method for computing a good basis.

### 3.2 Selecting the best basis

A complete binary tree provides a *dictionary* of more than one orthogonal basis from which the optimal basis can be adaptively chosen to represent the signal. The  $l^1$  cost of representing a length- $N$  signal  $x$  in the basis  $B = \{b_m\}$  is

$$C(x, B) = \sum_{m=1}^N \frac{|\langle x, b_m \rangle|}{\|x\|} \quad (9)$$

which provides a convenient measure of sparsity [1]. The *best basis* is the one which minimises this cost. The computationally efficient Coifman-Wickerhauser algorithm takes advantage of the binary structure and determines the best basis in  $O(N(\log_2 N)^2)$  time [2].

### 3.3 Adapting to the input

Our first method for adapting the basis attempts to maximise the sparsity of the average of the two mixtures

$$x_a = \frac{1}{2}(x_1 + x_2) \quad (10)$$

by minimising the  $l^1$  cost described in Section 3.2. This method will be referred to as *CPI*. Results for CPI are given in Section 4.

### 3.4 Adapting to a single source

One issue with the CP1 method is that it models mixtures of the sources rather than the sources themselves. To overcome this limitation, we adapt one basis to each source with the intention that each such basis will capture the time-frequency structures of that source better than the basis determined by CP1.

A heuristically motivated cost function maximises the energy of the local cosine coefficients associated with a particular source angle  $\theta_j$ . By selecting a basis which maximises the energy of coefficients that cluster around  $\theta_j$  we would expect that a sparse representation will be generated. The cost function is defined by

$$C(x_1, x_2, B, \theta_j, u) = - \sum_{m=1}^N \left( \Lambda_{\theta_j, u} \cdot \langle (x_1 \cos \theta_j + x_2 \sin \theta_j), b_m \rangle^2 \right) \quad (11)$$

where

$$\Lambda_{\theta_j, u} = \begin{cases} 1 & \text{if } \theta_j - \frac{u}{2} < \arctan \left( \frac{\langle x_2, b_m \rangle}{\langle x_1, b_m \rangle} \right) < \theta_j + \frac{u}{2} \\ 0 & \text{otherwise} \end{cases} \quad (12)$$

and  $B = \{b_m\}$  is a basis from the dictionary of bases derived from the complete local cosine tree. The function  $\Lambda_{\theta_j, u}$  masks local cosine coefficients in the basis  $B$ . The Coifman-Wickerhauser algorithm [2] finds the best basis corresponding to this cost function. This method will be referred to as *CP2*. The CP2-based technique learns a basis from a dictionary of bases; however, its advantage over techniques based on, for example, the matching pursuit algorithm [7] stems from the applicability of the fast tree-searching algorithm.

## 4 EVALUATION

We obtained eight pieces of multitracked music by several artists with access to the original multitracked digital audio data. (For details on these sources and more extensive evaluation, see [8].) This provided us with sources from which to synthesise instantaneous mixtures. For each mixture, the pitched sources were harmonically related, so overlapping partial frequencies were expected. Each source had a sample rate of 22.05 kHz at a resolution of 16 bits per sample. An extract of  $2^{18}$  samples was taken from each source, giving approximately 11.9 s of audio.

For each experiment, the mixtures  $x_1$  and  $x_2$  were generated with the same instantaneous mixing parameters:

$$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 0.83 & 0.60 & 0.40 & 0.83 \\ 0.17 & 0.40 & 0.60 & 0.17 \end{pmatrix} \begin{pmatrix} s_1 \\ s_2 \\ s_3 \\ s_4 \end{pmatrix}. \quad (13)$$

The resulting mixtures are simulations of *pan-potted stereo* mixing.

src	trans	K	D	u	SDR	SIR	SAR
$s_1$	STFT	16384	-	0.30	<b>21.93</b>	<b>38.97</b>	<b>22.02</b>
	MDCT	16384	-	0.20	18.03	34.48	18.14
	CP1	2048	7	0.20	18.69	36.64	18.76
	CP2	8192	5	0.30	16.90	27.43	17.31
$s_2$	STFT	32768	-	0.39	<b>9.82</b>	<b>38.26</b>	<b>9.83</b>
	MDCT	8192	-	0.39	6.91	29.71	6.94
	CP1	8192	5	0.39	6.58	31.77	6.60
	CP2	1024	8	0.39	7.82	32.97	7.83
$s_3$	STFT	32768	-	0.10	<b>2.26</b>	<b>12.66</b>	<b>2.90</b>
	MDCT	16384	-	0.10	-0.42	12.17	0.08
	CP1	64	12	0.10	-0.20	11.90	0.34
	CP2	16384	4	0.10	-1.72	9.56	-0.93
$s_4$	STFT	16384	-	0.39	1.52	20.88	1.60
	MDCT	16384	-	0.39	2.82	20.44	2.94
	CP1	16384	4	0.39	2.82	20.44	2.94
	CP2	512	9	0.39	<b>3.49</b>	<b>22.29</b>	<b>3.57</b>

Table 1: Separation results for one mixture, representative of typical performance over all experiments. All numerical measurements are in dB.

### 4.1 Measuring performance

For each estimated source, we make numerical evaluations of the contribution of unwanted sources (interference), by measuring the *Source to Interference Ratio (SIR)*, and the distortion due to the separation process (artefacts), by measuring the *Source to Artefacts Ratio (SAR)*. In order to simplify direct comparisons, the *Source to Distortion Ratio (SDR)* is computed; this combines both the SIR and SAR into a single numerical measure of total relative distortion. Methods for computing these measurement criteria are explained in detail in [11].

### 4.2 Experiments

Time-frequency masks were constructed and applied to the mixture channels  $x_1$  and  $x_2$  represented by the following transforms. For the STFT (Hamming-windowed with half-block overlap) and MDCT transforms, block sizes  $K = 2^m$  with  $m = 6, 7, \dots, 15$  were tested. (The STFT is complex-valued, so the masks were determined from the magnitude of the transform.) The CP1 and CP2 methods were evaluated with maximum tree search depths  $D = 3, 4, \dots, 12$ . The lengths of the CP1 and CP2 basis functions at each depth  $d$  in a local cosine tree correspond to block sizes  $K = 2^{18-d}$  (the length each input signal is  $2^{18}$  samples); this allows us to compare methods based on the lengths of their analysis windows. Values of  $u$  tested were 0.1, 0.2, 0.3, and 0.39.

Representative results from one of the mixtures are presented in Table 1. They indicate that the effects of the artefacts dominate interference, since the SAR values are typically significantly lower than SIR. This is not too surprising since we would expect the masking process to introduce some artefacts.

The number of ‘best’ results for each transform are shown in Table 2. Overall, the STFT showed best separation performance for the majority of sources and pieces, with our proposed CP2 showing best performance on most of the remainder. Nevertheless, since the CP2 method is a complete, orthogonal transform while the STFT is an overcomplete, non-orthogonal transform, with double

trans.	SDR	SIR	SAR	total
STFT	24 (75%)	18 (56%)	22 (69%)	64 (67%)
MDCT	2 (6%)	5 (16%)	1 (3%)	8 (8%)
CP1	0 (0%)	6 (19%)	0 (0%)	6 (6%)
CP2	6 (19%)	3 (9%)	9 (28%)	18 (19%)
total	32 (100%)	32 (100%)	32 (100%)	96 (100%)

Table 2: Performance of the various transforms. Each cell in the table indicates the number of times the transform scored best for that performance measure.

the representation size of CP2, we consider these results to represent competitive performance. We could find no immediately apparent relationship between the nature of an extracted source and the performances of the various transforms.

Informal listening tests indicate that in general our proposed CP2 method, when giving reasonable performance, appears to produce less audible ‘pipe noise’ compared to the STFT. These also suggest that the noise is least objectionable for mid-range tree depths (around  $D = 6$ ,  $K = 4096$ ), with more pipe noise for deeper trees (large  $D$ , small  $K$ ), while shallower trees (small  $D$ , large  $K$ ) are associated with pre-echo and apparent note timing jitter. We intend to carry out other listening tests in future to further investigate these effects.

## 5 FURTHER WORK

Results have shown that adapting a local cosine basis to the output can give good results. However, the energy-based cost function (Equation (11)) is derived from heuristic reasoning. It may be the case that a more subtle cost function is required to represent the estimated source more sparsely. In particular, the current energy-based cost function considers only coefficients of the estimated source without regarding the coefficients of the other sources.

Finally, the performance measures, SIR, SAR and SDR, may not correspond well to a subjective human assessment of separation performance. Informal listening tests show that each representation imparts a noticeably different timbre to the extracted sources. Therefore, we believe that listening tests would give a more meaningful, practical measure of separation performance.

## 6 CONCLUSIONS

We have described a time-frequency masking approach to stereo audio source separation using local cosine packet representations. The CP2 method adapts the basis to the time-frequency mask for each separate basis, is fast and gives promising results.

We compared the performance of our proposed methods to the STFT and the MDCT on a set of instantaneous stereo musical audio mixtures. The STFT gives the best performance for separation of most sources from most mixtures. Nevertheless, our results indicate that the performance our proposed CP2 method is competitive, and exhibits better performance than the MDCT. Informal listening tests suggest that the cosine packet method can exhibit less objectionable noise than the STFT.

## ACKNOWLEDGEMENTS

Andrew Nesbit is supported by the Department of Electronic Engineering, Queen Mary, University of London and by European Commission grant FP6-IST-507142. This work is also partially supported by EPSRC grants GR/S75802/01 and GR/S85900/01.

## References

- [1] S. S. Chen, D. L. Donoho, and M. A. Saunders. Atomic decomposition by basis pursuit. *SIAM Review*, 43(1):129–159, 2001.
- [2] R. R. Coifman and M. V. Wickerhauser. Entropy-based algorithms for best basis selection. *IEEE Transactions on Information Theory*, 38(2):713–718, Mar. 1992.
- [3] 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.
- [4] P. Georgiev, F. Theis, and A. Cichocki. Sparse component analysis and blind source separation of underdetermined mixtures. *IEEE Transactions on Neural Networks*, 16(4):992–996, July 2005.
- [5] R. Gribonval. Piecewise linear source separation. In M. A. Unser, A. Aldroubi, and A. F. Laine, editors, *Proceedings of the SPIE (Wavelets: Applications in Signal and Image Processing X)*, volume 5207, pages 297–310. SPIE—The International Society for Optical Engineering, WA, USA, Nov. 2003.
- [6] P. Kisilev, M. Zibulevsky, Y. Y. Zeevi, and B. A. Pearlmutter. Multiresolution framework for blind source separation. Technical Report CCIT 317, Technion University, Israel, June 2001.
- [7] S. Mallat. *A Wavelet Tour of Signal Processing*. Academic Press, second edition, 1999.
- [8] A. Nesbit, M. D. Plumbley, and M. E. Davies. Audio source separation by time-frequency masking using a signal-adaptive local cosine transform. Submitted for publication, 2006.
- [9] P. D. O’Grady, B. A. Pearlmutter, and S. T. Rickard. Survey of sparse and non-sparse methods in source separation. *International Journal of Imaging Systems and Technology (Special Issue: Blind Source Separation and Deconvolution in Imaging and Image Processing)*, 15(1):18–33, 2005.
- [10] V. Y. F. Tan and C. Févotte. A study of the effect of source sparsity for various transforms on blind audio source separation performance. In *Proceedings of the Workshop on Signal Processing with Adaptive Sparse Structured Representations (SPARS’05)*, Rennes, France, 16–18 November 2005.
- [11] E. Vincent, R. Gribonval, and C. Févotte. Performance measurement in blind audio source separation. *IEEE Transactions on Speech and Audio Processing*, 2004. Preprint, to appear.
- [12] E. Vincent, M. G. Jafari, S. A. Abdallah, M. D. Plumbley, and M. E. Davies. Blind audio source separation. Technical Report C4DM-TR-05-01, Centre for Digital Music, Queen Mary, University of London, Nov. 2005. Available online <http://www.elec.qmul.ac.uk/people/emmanuelv/VincentEtAl05.bass-tutorial.pdf>.
- [13] Ö. Yılmaz and S. Rickard. Blind separation of speech mixtures via time-frequency masking. *IEEE Transactions on Signal Processing*, 52(7):1830–1847, July 2004.