# A STUDY OF THE EFFECT OF SOURCE SPARSITY FOR VARIOUS TRANSFORMS ON BLIND AUDIO SOURCE SEPARATION PERFORMANCE

*Vincent Y. F. Tan*[*]

Human Factors Lab
DSO National Laboratories
Singapore 117510
tyanfu@dso.org.sg

*Cédric Févotte*[†]

Engineering Department
University of Cambridge
Cambridge CB2 1PZ, UK
cf269@eng.cam.ac.uk

## ABSTRACT

In this paper, the problem of blind separation of underdetermined noisy mixtures of audio sources is considered. The sources are assumed to be sparsely represented in a transform domain. The sparsity of their analysis coefficients is modelled by the Student $t$ distribution. This prior allows for robust Bayesian estimation of the sources, the mixing matrix, the additive noise variance as well as hyperparameters of the Student $t$ priors, using a Gibbs sampler (a standard Monte Carlo Markov Chain simulation method). The performances resulting from the use of various transforms, orthonormal as well as overcomplete, are compared. More precisely, we present extensive separation results of $2 \times 3$ mixtures of various sets of sources (speech, musical and percussion sources) using various orthonormal transforms - Discrete Cosine Transform (DCT), Modified Discrete Cosine Transform (MDCT), Discrete Wavelet Transforms (DWT), and overcomplete transforms - Short-Time Discrete Cosine Transform (STDCT) and union of MDCT and DWT. In general, the MDCT is found to be a good transform to use on the various types of audio signals. However, the DWT is the best transform to use on the percussion signals as they contain more transients than tonals. The provided results show that the separation performance is indeed correlated to the sparsity of the analysis coefficients of the sources in the transform domain. Our results also show that the use of overcomplete transforms does not lead to significant improvement in performance, because they fail to improve the sparsity measure.

## 1. INTRODUCTION

Blind Source Separation (BSS) consists of estimating $n$ signals (the sources) from the sole observation of $m$ mixtures of them (the observations). In this paper we consider linear instantaneous mixtures of time series: at each time index, the observations are a linear combination of the sources at the same time index. The (over)determined problem ($m \geq n$) has been widely studied, in particular within the field of Independent Component Analysis (see [1] for an overview). In this paper, we consider the underdetermined case. This case is very challenging because contrary to (over)determined mixtures, estimating the mixing system is not sufficient for reconstruction of the sources, since for $m < n$ the mixing matrix is not invertible. Then, it appears that separation of underdetermined mixtures requires important prior information on the sources to allow their reconstruction. Prior information is also helpful for reconstructing the sources in noisy environments.

A now common approach to BSS, in particular for mixtures possibly underdetermined and noisy, is the use of source sparsity assumptions (see [2]). The assumption of sparsity means that only a few coefficients of the sources are significantly non-zero. If the sources are not sparse in their original domain (e.g, the time domain for audio signals), they might be sparse in a transformed domain (e.g, the Fourier domain, wavelet transform).

In this work we are interested in studying the influence of the choice of the transform over audio source separation quality. We applied a source separation algorithm to the analysis coefficients of the observations of a noisy $2 \times 3$ mixture with 4 different sets of sources. The first set is composed of speech sources, the second set of musical sources, the third set of percussion sources and the fourth set is composed of a speech source, a musical source and a percussion source. Various real-valued orthonormal transforms are considered, namely the Discrete Cosine Transform (DCT), the Modified Discrete Cosine Transform (MDCT), Discrete Wavelet Transforms (DWT), as well as overcomplete transforms, namely the Short-Time Discrete Cosine Transform (STDCT) and an hybrid transform (MDCT + DWT).

The separation algorithm applied to the observation analysis coefficients (that is, the coefficients yielded by the transform applied to each of the observations) is described in [3]. It is a Bayesian method in which the source analysis coefficients are given a Student $t$ prior, which leads to a sparse prior for low degrees of freedom. A Gibbs sampler is used to sample from the posterior distribution of the sources, the mixing matrix, the additive noise variance as well as hyperparameters of the Student $t$ priors.

In general, the MDCT is found to be a good transform to use on the various types of audio signals. However, the DWT is the best transform to use on the percussion signals as they contain more transients than tonals. The provided results show that separation performance is correlated to the sparsity of the analysis coefficients of the sources in the transform domain. Our results also show that the use of overcomplete transforms does not lead to significant improvement in performance, because they fail to improve the sparsity measure of the source coefficients.

This paper is organized as follows: in Section 2 we introduce the linear instantaneous noisy model and the source and noise assumptions, in Section 3 we describe briefly the separation scheme, Section 4 describes the experimental framework and Section 5 gives results. Conclusions and perspectives are presented in Section 6.

[*] Vincent Tan performed this piece of research under Dr. Févotte's supervision, as a partial fulfilment of his M.Eng degree at the Engineering Department of the University of Cambridge.

[†] Cédric Févotte acknowledges support from the European Commission funded Research Training Network HASSIP (HPRN-CT-2002-00285).

## 2. MODEL AND ASSUMPTIONS

### 2.1. Mixture model

We consider the following standard linear instantaneous model:

$$\mathbf{x}_t = \mathbf{A}\mathbf{s}_t + \mathbf{n}_t, \qquad 1 \le t \le N \tag{1}$$

where $\mathbf{x}_t = (x_{1,t}, \ldots, x_{m,t})^T$ is the vector of size $m$ containing the observations, $\mathbf{s}_t = (s_{1,t}, \ldots, s_{n,t})^T$ is the vector of size $n$ containing the sources and $\mathbf{n}_t$ is an additive noise vector. $\mathbf{A}$ is the $m \times n$ unknown full rank mixing matrix (with possibly $m < n$). Variables without time index $t$ denote whole sequences or samples, for example, $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_N)$. As a consequence, (1) can be rewritten more succinctly as

$$\mathbf{x} = \mathbf{A}\mathbf{s} + \mathbf{n}. \tag{2}$$

The goal is to estimate the sources $\mathbf{s}$ and the mixing matrix $\mathbf{A}$ up to the standard BSS indeterminacies on gain and order [1].

### 2.2. Assumptions

*2.2.1. Transform domain*

Given an analysis operator $\mathbf{\Psi} \in \mathbb{R}^{N\times K}$ corresponding to a chosen linear real-valued transform, the sequence of analysis coefficients of a signal $x \in \mathbb{R}^{1\times N}$ is the sequence $\tilde{x} \in \mathbb{R}^{1\times K}$ given by

$$\tilde{x} = x\,\mathbf{\Psi} \tag{3}$$

with possibly $K > N$ (overcomplete transform). Denoting $\tilde{\mathbf{x}} = \tilde{\mathbf{x}}\,\mathbf{\Psi}$ and $\tilde{\mathbf{s}}$ and $\tilde{\mathbf{n}}$ accordingly, in the transform domain equation (2) reads

$$\tilde{\mathbf{x}} = \mathbf{A}\tilde{\mathbf{s}} + \tilde{\mathbf{n}}. \tag{4}$$

*2.2.2. Model of the sources*

The underlying assumption of the following work is that the sequences $\tilde{s}_1, \ldots, \tilde{s}_n$ are sparse. Following the work in [3] the source analysis coefficients are given a Student $t$ prior which gathers most of its probability mass around zero and exhibits heavy tails for small degrees of freedom, and is thus a relevant model of sparsity (see details in [3]). Each sequence $\tilde{s}_i$ is modeled as an identically and independently distributed sequence, and the sequences $\tilde{s}_1, \ldots, \tilde{s}_n$ are modeled as mutually independent.

*2.2.3. Noise assumptions*

We assume that $\tilde{\mathbf{n}}$ is a i.i.d Gaussian noise with covariance $\sigma^2\,\mathbf{I}_m$, and with $\sigma$ unknown. We point out that when an orthonormal basis is used (*i.e*, $K = N$ and $\mathbf{\Psi}^{-1} = \mathbf{\Psi}^T$), $\mathbf{n}$ is equivalently a Gaussian i.i.d noise with covariance $\sigma^2\,\mathbf{I}_m$.

## 3. METHOD

The Gibbs sampler presented in [3] allows to decompose the data $\tilde{\mathbf{x}}$ as:

$$\tilde{\mathbf{x}} = \widehat{\mathbf{A}}\,\widehat{\tilde{\mathbf{s}}} + \widehat{\tilde{\mathbf{n}}} \tag{5}$$

where $\widehat{\mathbf{A}}$ and $\widehat{\tilde{\mathbf{s}}}$ are Minimum Mean Squared Estimates (MMSE) of $\mathbf{A}$ and $\tilde{\mathbf{s}}$. Time domain estimates of the sources can be obtained by applying the synthesis operator $\mathbf{\Phi} = \mathbf{\Psi}^T(\mathbf{\Psi}\,\mathbf{\Psi}^T)^{-1} \in \mathbb{R}^{K\times N}$ to $\widehat{\tilde{\mathbf{s}}}$:

$$\widehat{\mathbf{s}} = \widehat{\tilde{\mathbf{s}}}\mathbf{\Phi} \tag{6}$$

Note that when $K = N$, $\mathbf{\Psi} = \mathbf{\Phi}^{-1}$.

## 4. EXPERIMENTAL FRAMEWORK

We present separation results of $2 \times 3$ mixtures of various sets of sources using various transforms. The sources, mixtures and estimates obtained with all the transforms for each data set can be listened to at `http://www-sigproc.eng.cam.ac.uk/~cf269/spars05/sound_files.html`.

### 4.1. Audio data

The sets of sources are: speech sources $\{s_1^{sp}, s_2^{sp}, s_3^{sp}\}$, musical sources $\{s_1^{mu}, s_2^{mu}, s_3^{mu}\}$, percussion sources $\{s_1^{pe}, s_2^{pe}, s_3^{pe}\}$, and the combination $\{s_1^{sp}, s_2^{mu}, s_3^{pe}\}$. The signals are sampled at 8kHz, with $N = 65536$ (approx. 8s). The same mixing matrix was used to generate synthetic mixtures of the above sets of sources, and we used:

$$\mathbf{A} = \begin{bmatrix} 0.8\cos(-\pi/3) & 0.9\cos(-\pi/8) & 0.8\cos(\pi/4) \\ 0.8\sin(-\pi/3) & 0.9\sin(-\pi/8) & 0.8\sin(\pi/4) \end{bmatrix} \tag{7}$$

$\approx$ 16dB i.i.d Gaussian noise was added to the observations (*in the time domain*).

### 4.2. Transforms

We used orthonormal as well as overcomplete transforms. Because the method [3] is implemented for real-valued data, we used real-valued transforms. See [4] for an overview of the transforms used. The orthonormal bases are:

**1 - DCT** Discrete Cosine Transform,

**2 - MDCT** Modified Discrete Cosine Transform (a local cosine transform) used with a sine bell window and 50% of overlap, and with a time resolution of 64ms (half the window length),

**3 - DWT-Vai** Discrete Wavelet Transform with Vaidyanathan filter (with good properties for speech coding),

**4 - DWT-Sym** Discrete Wavelet Transform with Symmlet filter of order $p = 8$,

**5 - WPBB** Wavelet Packet Best Basis applied to $x_1$ with minimum $l_1$ norm criterion,

**6 - NT** No transform (the time samples are processed).

The overcomplete transforms are:

**1 - STDCT** Short-Time Discrete Cosine Transform (with a window length of 128ms and 25% overlap),

**2 - HT** Hybrid Transform: union of MDCT and DWT transforms.

### 4.3. Sparsity measure

To quantify sparsity, we use the sparsity index for a signal vector $\tilde{s}_i$ defined by

$$\xi \triangleq \frac{\|\tilde{s}_i\|_1}{\|\tilde{s}_i\|_2}. \tag{8}$$

The smaller $\xi$ is the sparser the signal vector $\tilde{s}_i$.

### 4.4. Numerical criteria for source separation evaluation

We used the criteria described in [5]. Basically, the SDR (Source to Distortion Ratio) provides an overall separation performance criterion, the SIR (Source to Interferences Ratio) measures the level of interferences from the other sources in each source estimate, SNR (Source to Noise Ratio) measures the error due to the additive noise

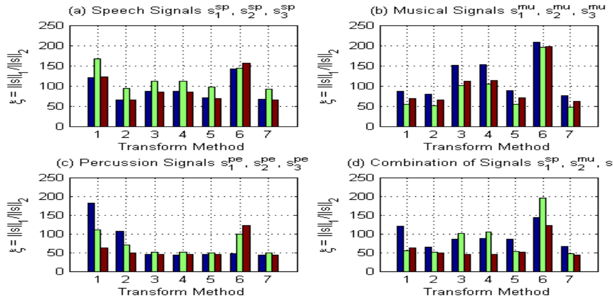

**Fig. 1:** Sparsity Indices for Sources using Orthonormal Bases a) Speech Sources, b) Musical Sources, c) Percussion Sources, d) Combination of Sources; Transform method: 1-DCT, 2-MDCT, 3-WT-Vai, 4-WT-Sym, 5-WPBB, 6-NT, 7-WPBB on sources (optimal)

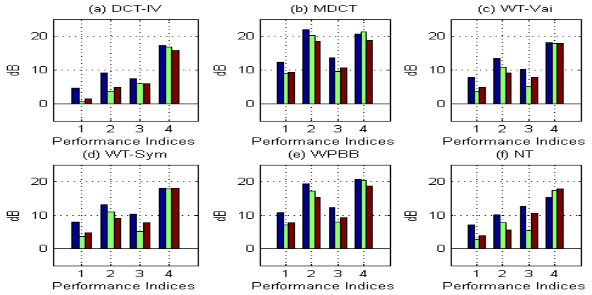

**Fig. 2:** Performances of the various transforms on **Speech Sources**; Performance Indices: 1-SDR, 2-SIR, 3-SAR, 4-SNR; Sources: blue-$s_1^{sp}$, green-$s_2^{sp}$, brown-$s_3^{sp}$

on the sensors and the SAR (Source to Artifacts Ratio) measures the level of artifacts in the source estimates. The higher are the ratios, the better is quality of estimation.

## 5. SEPARATION RESULTS

### 5.1. Orthonormal transforms

Fig. 1 gives the values of the sparsity indices of the analysis coefficients sequences obtained by applying the above orthonormal transforms to the 4 sets of sources. In addition to the above transforms we also give the sparsity index of the coefficients obtained from the best basis algorithm applied to *each source*. It hence give an optimal value of $\xi$ for each of the source. Fig. 2, 3, 4 and 5 give the SDR, SIR, SAR and SNR of the source estimates provided by the source separation algorithm applied to the analysis coefficients of the data, obtained from the various orthonormal transforms and using the various sets of data. Fig. 1 shows that sparsest representations of speech signals are obtained with MDCT or WPBB applied to $x_1$, and Fig. 2 shows that the best separation results are obtained with these transforms too. Sparsest representations of music signals were obtained with either the DCT, the MDCT or the WPBB, which led to best separation results too. Sparsest representations of percussions signals were obtained using DWTs, which led again to the best separation results. MDCT performed rather well too. Fig. 5 shows that, when dealing with a mixture of sources of various types, best results are obtained with the MDCT.

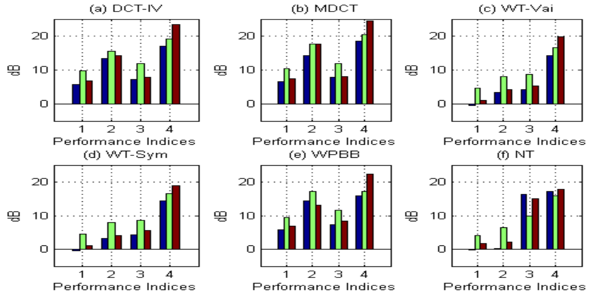

**Fig. 3:** Performances of the various transforms on **Musical Sources**; Performance Indices: 1-SDR, 2-SIR, 3-SAR, 4-SNR; Sources: blue-$s_1^{mu}$, green-$s_2^{mu}$, brown-$s_3^{mu}$

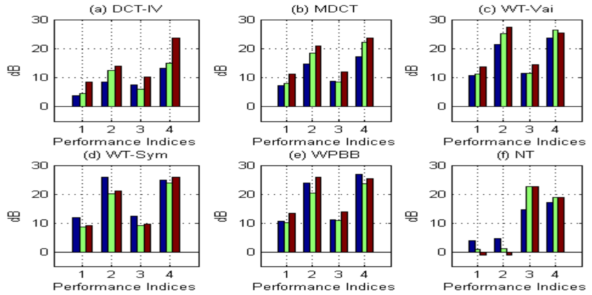

**Fig. 4:** Performances of the various transforms on **Percussion Sources**; Performance Indices: 1-SDR, 2-SIR, 3-SAR, 4-SNR; Sources: blue-$s_1^{pe}$, green-$s_2^{pe}$, brown-$s_3^{pe}$

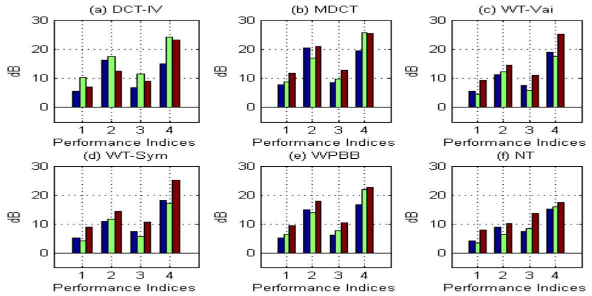

**Fig. 5:** Performances of the various transforms on **Combination of Sources**; Performance Indices: 1-SDR, 2-SIR, 3-SAR, 4-SNR; Sources: blue-$s_1^{sp}$, green-$s_2^{mu}$, brown-$s_3^{pe}$

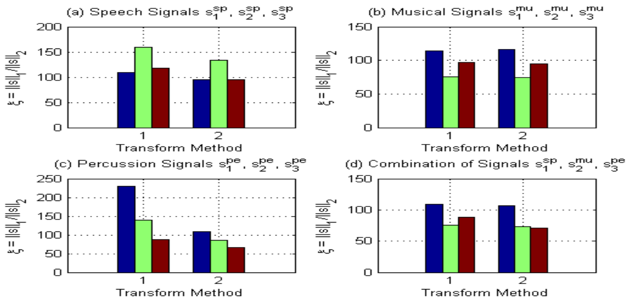

**Fig. 6:** Sparsity Indices for Sources using Overcomplete Dictionaries a) Speech Sources, b) Musical Sources, c) Percussion Sources, d) Combination of Sources; Transform method: 1-STDCT, 2-HT

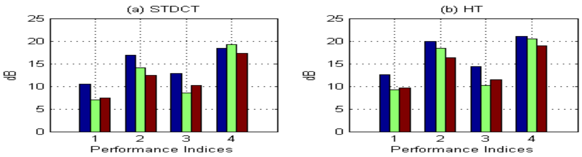

**Fig. 7:** Performances of the two Overcomplete Dictionaries on **Speech Sources**; Performance Indices: 1-SDR, 2-SIR, 3-SAR, 4-SNR; Sources: blue-$s_1^{sp}$, green-$s_2^{sp}$, brown-$s_3^{sp}$

### 5.2. Overcomplete transforms

Fig. 6 gives the values of the sparsity indices of the analysis coefficients sequences obtained by applying the above overcomplete transforms to the 4 sets of sources. Fig. 7, 8, 9 and 10 give the SDR, SIR, SAR and SNR of the source estimates provided by the source separation algorithm applied to the analysis coefficients of the data.

Fig. 6 shows that the overcomplete transforms fail to improve the sparsity of the source coefficients, when compared with the orthonormal transforms. The separation performance obtained with the various sets of sources does not increase.

## 6. CONCLUSIONS

This work leads to four conclusions:

- the MDCT basis yields good separation results for audio signals of various types,
- it is a good idea to apply the WPBB algorithm to one of the observations, this can lead to better results than those obtained with the MDCT when the sources belong to the same class of signals (in particular for percussion sources)

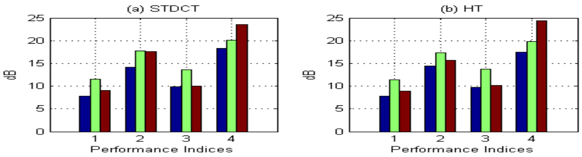

**Fig. 8:** Performances of the two Overcomplete Dictionaries on **Musical Sources**; Performance Indices: 1-SDR, 2-SIR, 3-SAR, 4-SNR; Sources: blue-$s_1^{mu}$, green-$s_2^{mu}$, brown-$s_3^{mu}$

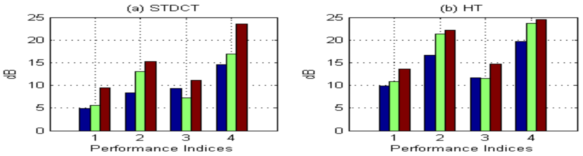

**Fig. 9:** Performances of the two Overcomplete Dictionaries on **Percussion Sources**; Performance Indices: 1-SDR, 2-SIR, 3-SAR, 4-SNR; Sources: blue-$s_1^{pe}$, green-$s_2^{pe}$, brown-$s_3^{pe}$

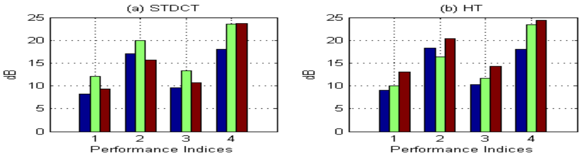

**Fig. 10:** Performances of the two Overcomplete Dictionaries on the **Combination of Sources**; Performance Indices: 1-SDR, 2-SIR, 3-SAR, 4-SNR; Sources: blue-$s_1^{sp}$, green-$s_2^{mu}$, brown-$s_3^{pe}$

- our experimental results tend to show that the separation quality is correlated with the sparsity of the source coefficients - however a rigorous proof of this observation is still to be given,
- it does not appear to be necessary to use overcomplete transforms as they fail to improve the sparsity of the source analysis coefficients and do not improve the separation results. Note however that it does not mean than overcomplete dictionaries are useless to source separation. Indeed, a good strategy would be to model the sources as sparse linear combinations of atoms in an overcomplete dictionary, and thus try to estimate the synthesis coefficients instead of the analysis coefficients. These models lead to more complicated separation schemes that are still to be developed.

## 7. REFERENCES

[1] Hyvärinen, A., Karhunen, J. and Oja, E., *Independent Component Analysis*, 1st ed. Wiley Interscience, 2001.

[2] M. Zibulevsky, B. A. Pearlmutter, P. Bofill, and P. Kisilev, “Blind source separation by sparse decomposition,” in *Independent Component Analysis: Principles and Practice*, S. J. Roberts and R. M. Everson, Eds. Cambridge University Press, 2001.

[3] C. Févotte and S. Godsill, “A Bayesian approach to blind separation of sparse sources,” *IEEE Transactions on Speech and Audio Processing*, 2005, in press. Preprint available at http://www-sigproc.eng.cam.ac.uk/~cf269/Journals/ieee_sap06.pdf.

[4] Mallat, S., *A Wavelet Tour of Signal Processing*. San Diego, CA: Academic Press, 1999.

[5] R. Gribonval, L. Benaroya, E. Vincent, and C. Févotte, “Proposals for performance measurement in source separation,” in *Proc. 4th Symposium on Independent Component Analysis and Blind Source Separation (ICA’03)*, Nara, Japan, Apr. 2003.