Sparse Representations in Audio and Music: from Coding to Source Separation

M. D. Plumbley, *Member, IEEE,* T. Blumensath, *Member, IEEE,* L. Daudet, *Member, IEEE,* R. Gribonval, *Senior Member, IEEE,* and M. E. Davies, *Member, IEEE*

***Abstract*—Sparse representations have proved a powerful tool in the analysis and processing of audio signals and already lie at the heart of popular coding standards such as MP3 and Dolby AAC. In this paper we give an overview of a number of current and emerging applications of sparse representations in areas from audio coding, audio enhancement and music transcription to blind source separation solutions that can solve the "cocktail party problem". In each case we will show how the prior assumption that the audio signals are approximately sparse in some time-frequency representation allows us to address the associated signal processing task.**

I. Introduction

Over recent years there has been growing interest in finding ways to transform signals into *sparse representations*, i.e. representations where most coefficients are zero. These sparse representations are proving to be a particularly interesting and powerful tool for analysis and processing of audio signals.

Audio signals are typically generated either by resonant systems or by physical impacts, or both. Resonant systems produce sounds that are dominated by a small number of frequency components, allowing a sparse representation of the signal in the frequency domain. Impacts produce sounds that are concentrated in time, allowing a sparse representation of the signal in either directly the time domain, or in terms of a small number of wavelets. The use of sparse representations therefore appears to be a very appropriate approach for audio.

Manuscript received January XX, 200X; revised January XX, 200X. This work was supported in part by the EU Framework 7 FET-Open project FP7-ICT-225913-SMALL: Sparse Models, Algorithms and Learning for Large-Scale data

M. D. Plumbley is with Queen Mary University of London, School of Electronic Engineering and Computer Science, Mile End Road, London E1 4NS, UK (e-mail: mark.plumbley@elec.qmul.ac.uk). He is supported by a Leadership Fellowship from the UK Engineering and Physical Sciences Research Council (EPSRC).

T. Blumensath is with the School of Mathematics, University of Southampton, Southampton, SO17 1BJ, UK (e-mail: thomas.blumensath@soton.ac.uk).

L. Daudet was at the time of writing with LAM, Institut Jean Le Rond dAlembert, Université Pierre et Marie Curie (UPMC Univ. Paris 06), 11 rue de Lourmel, 75015 Paris, France (e-mail: daudet@lam.jussieu.fr). In September 2009 he joined the Langevin Institute for Waves and Images (LOA), Université Denis DiderotParis 7.

R. Gribonval is with INRIA, Centre Inria Rennes—Bretagne Atlantique, Campus de Beaulieu, F-35042 Rennes Cedex, Rennes, France (e-mail: remi.gribonval@inria.fr).

M. E. Davies is with the Institute for Digital Communications (IDCOM) & Joint Research Institute for Signal and Image Processing, School of Engineering and Electronics, University of Edinburgh, The King's Buildings, Mayfield Road, Edinburgh EH9 3JL, Scotland, UK (e-mail: mike.davies@ed.ac.uk). He acknowledges support of his position from the Scottish Funding Council and their support of the Joint Research Institute in Signal and Image Processing with the Heriot-Watt University as a component part of the Edinburgh Research Partnership.

In this article, we will examine a range of applications of sparse representations to audio and music signals. We will see how this concept of sparsity can be used to design new methods for audio coding which have improved performance over non-sparse methods; how it can be used to perform denoising and enhancement on degraded audio signals; and how it can be used to separate source signals from mixed audio signals, particularly when there are more sources than microphones. Finally, we will also see how finding a sparse decomposition can lead to a note-like representation of musical signals, similar to automatic music transcription.

A. Sparse Representations of an Audio Signal

Suppose we have a sampled audio signal with T samples $x(t)$, $1 \le t \le T$, which we can write in a row vector form as $\bar{x} = (x(1), \ldots, x(T))$. For audio signals we are typically dealing with signals sampled below 20 kHz, but for simplicity we will assume our sampled time t takes integer values. It is often convenient to decompose $\bar{x}$ into a weighted sum of Q basis vectors $\bar{\phi}_q = (\phi_q(1), \ldots, \phi_q(T))$, with the contribution of the q-th basis vector weighted by a coefficient u_q:

$$x(t) = \sum_{q=1}^{Q} u_q \phi_q(t) \qquad \text{or} \qquad \bar{x} = \sum_{q=1}^{Q} u_q \bar{\phi}_q \tag{1}$$

or in matrix form

$$\bar{x} = \bar{u}\mathbf{\Phi} \tag{2}$$

where $\mathbf{\Phi}$ is the matrix with elements $[\mathbf{\Phi}]_{qt} = \phi_q(t)$.

The most familiar representation of this type in audio signal processing is the (Discrete) *Fourier* representation. Here we have the same number of basis vectors as signal samples ($Q = T$), and the basis matrix elements are given by

$$\phi_q(t) = \frac{1}{T} \exp\left(\frac{2\pi j}{T} qt\right) \tag{3}$$

where $j = \sqrt{-1}$. Now it remains for us to find the coefficients u_q in this representation of $\bar{x}$. In the case of our Fourier representation, this is straightforward: the matrix $\mathbf{\Phi}$ is *square and invertible*, and in fact orthogonal, so $\bar{u}$ can be calculated directly as $\bar{u} = \bar{x}\mathbf{\Phi}^{-1} = \bar{x}(T\mathbf{\Phi}^{\mathrm{H}})$, where the superscript $^{\mathrm{H}}$ denotes the conjugate transpose.

Signal representations corresponding to invertible transforms such as the DFT, the discrete cosine transform (DCT), or the discrete wavelets transform (DWT) are convenient and easy to calculate. However, it is possible to find many alternative representations. In particular, if we allow the number of

basis vectors (and hence coefficients) to exceed the number of signal samples, $Q > T$, then solving (2) for the representation coefficient vector $\bar{u}$ is in general not unique: there will be a whole $(Q-T)$-dimensional subspace of vectors $\bar{u}$ which satisfy $\bar{x} = \bar{u}\mathbf{\Phi}$. In this case we say that (2) is *underdetermined*. A common choice in this situation is to use the *Moore-Penrose pseudoinverse* $\mathbf{\Phi}^\dagger$, yielding $\bar{u} = \bar{x}\mathbf{\Phi}^\dagger$. However, in this article we are interested in finding representations that are *sparse*, i.e. representations where only a small number of the coefficients of $\bar{u}$ are non-zero.

B. Advantages of sparse representations

Finding a sparse representation for a signal has many advantages for applications such as coding, enhancement, or source separation. In coding, a sparse representation has only a few non-zero values, so only these values (and their locations) need to be encoded to transmit or store the signal. In enhancement, the noise or other disturbing signal is typically not represented by the same coefficients as the sparse signal. Therefore discarding the "wrong" coefficients can remove a large proportion of the unwanted noise, leaving a much cleaner restored signal. Finally, in source separation, if each signal to be separated has a sparse representation, then there is a good chance that there will be little overlap between the small sets of coefficients used to represent the different source signals. Therefore by selecting the coefficients "used" by each source signal, we can restore each of the original signals with most of the interference from the unwanted signals removed.

For typical steady-state audio signals, the Fourier representation already does quite a good job of providing an approximately sparse representation. If an audio signal consists of only stationary oscillations, without onsets or transients, a representation based on a short-time Fourier transform (STFT) or a Modified Discrete Cosine Transform (MDCT) [1] will include some large-amplitude coefficients corresponding to the main frequencies of oscillation of the signal, with little energy in between these.

However, audio signals also typically contain short transients at the onsets of musical notes or other sounds. These would not have a sparse representation in an STFT or MDCT basis, but instead in such a representation would require a large number of frequency components to be active simultaneously.

One approach to overcome this is therefore to look for a representation in terms of a *union of bases*, each with different time-frequency characteristics. For example, we could create a "tall, thin" basis matrix

$$\mathbf{\Phi} = \begin{bmatrix} \mathbf{\Phi}_{\mathrm{C}} \\ \mathbf{\Phi}_{\mathrm{W}} \end{bmatrix} \tag{4}$$

composed of both an MDCT basis $\mathbf{\Phi}_{\mathrm{C}}$, designed to represent the steady-state sinusoidal parts, and a Wavelet basis $\mathbf{\Phi}_{\mathrm{W}}$ designed to represent the transient, edge-like parts. We could write this representation as

$$\bar{x} = \bar{u}\mathbf{\Phi} = \bar{u}_{\mathrm{C}}\mathbf{\Phi}_{\mathrm{C}} + \bar{u}_{\mathrm{W}}\mathbf{\Phi}_{\mathrm{W}} \tag{5}$$

where the joint coefficient vector $\bar{u} = (\bar{u}_{\mathrm{C}}\ \bar{u}_{\mathrm{W}})$ is a concatenation of the MDCT and Wavelet coefficients. This type of idea is known in audio processing as *hybrid representations* [2] and also appears in image processing as *multilayered representations* [3] or *Morphological Component Analysis* [4]. Many other unions are possible, such as unions of MDCT bases with differing time-frequency resolutions. While the number of basis vectors, and hence the number of possible coefficients, is larger in a union of bases, we may find that the resulting representation has fewer non-zero coefficients and is therefore sparser (Fig. 1).

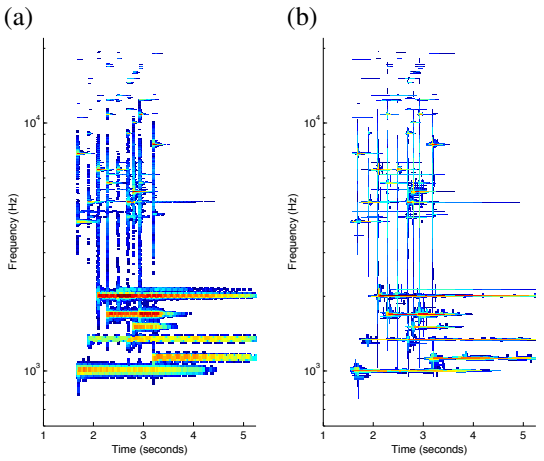

Fig. 1. Representations of an audio signal in (a) a single MDCT basis, and (b) a union of eight MDCT bases with different window sizes ("8*MDCT").

C. Recovering sparse representations

To find a sparse representation when the system is underdetermined is not quite so straightforward as for the square and invertible case. Finding the true sparsest representation

$$\arg\min_{\bar{u}} \{\|\bar{u}\|_0 \mid \bar{x} = \bar{u}\mathbf{\Phi}\} \tag{6}$$

where the 0-norm $\|\bar{u}\|_0$ is the number of non-zero elements of $\bar{u}$, is an NP-hard problem, so would take us a very long time to solve. However, it is possible to find an approximate solution to this. One method is to use the so-called *Basis Pursuit* relaxation, where instead of looking to solve (6) we look for a solution to the easier problem

$$\arg\min_{\bar{u}} \{\|\bar{u}\|_1 \mid \bar{x} = \bar{u}\mathbf{\Phi}\} \tag{7}$$

where the 1-norm $\|\bar{u}\|_1 = \sum_q |u_q|$ is the sum of the absolute values. Eqn. (7) is equivalent to a linear program (LP), and can be solved by a range of general or specialist methods, see e.g. [5], [6], [7], [8], [9], [10], [11].

Another alternative is to use a greedy algorithm to find an approximation to (6). For example, the matching pursuits (MP) algorithm [12] and orthogonal matching pursuit (OMP) algorithm [13] are well-known examples of this type of greedy algorithm. There are many more in the literature [14], [15], [16] and considerable recent work in the area of sparse representations has concentrated on theoretically optimal and practically efficient methods to find solutions (or approximate solutions) to (6) or (7). Nevertheless, MP is still used in

real-world problems since there are efficient implementations available, such as the Matching Pursuit Toolkit (MPTK)[1] [17].

II. Coding

Coding is arguably the most straightforward application of sparse representations. Indeed, reversibly transforming the signal into a new domain where the information is concentrated on a few terms is the main idea underlying data compression. The transform coder is a classical technique used in source coding [18]. When the transform is orthonormal it can be shown (under certain assumptions) that the gain achievable through transform coding is directly linked to the transform's ability to concentrate the energy of the signal in a small number of coefficients [19]. However, this problem is not as straightforward as it may seem, since there is no single fixed orthonormal basis where all audio signals have a sparse representation. Audio signals are in general quite diverse in nature: they mostly have a strong tonal part, but also some lower-energy components such as transients components (at note attacks) and wide-band noise that are nonetheless important in the perception of audio signals. These tonal, transient and noise components are optimally represented in bases with different respective requirements in terms of time-frequency localization.

We will consider two main approaches to handle this issue. The first approach is to find an *adapted* orthonormal basis, best fitted to the local features of the signal. This is the technique employed in most state-of-the-art commercial audio codecs, such as MPEG 2/4 Advanced Audio Codec (AAC). The second approach uses dictionary redundancy to accommodate this variety of features, leading to a sparser representation, but where each coefficient carries more information.

A. Coding in adapted orthogonal bases

For coding, using an orthonormal basis seems an obvious choice. Orthonormal bases yield invertible transforms with no redundancy, so the number of coefficients in the transform domain is equal to the number of samples. Many multimedia signals have compact representations in orthonormal bases: for example, images are often well suited to wavelet representations (EZW, JPEG200). Furthermore, several orthonormal schemes also have fast implementations due to the special structure of the basis, such as the FFT for implementing the DFT, or Mallat's multiresolution algorithm for the DWT [19].

For audio signals, a natural choice for an orthonormal transform might be to use one based on the STFT. However, for real signals the Balian-Low theorem tells us that there cannot be a real orthonormal transform based on local Fourier transforms with nice regularities properties both in time and frequency.

To overcome this we can use so-called *Lapped Orthogonal Transforms*, which exploit special aliasing cancellation properties of the cosine transform, when the window obeys two conditions on symmetry and energy-preservation. The discrete version of these classes of transforms leads to the Modified Discrete Cosine Transform (MDCT) [1], with atoms such as

$$\bar{\phi}_{k,p}(t) = h(\tau)\sqrt{\frac{2}{L}}\cos\left[\frac{\pi}{L}\left(\tau + \frac{1+L}{2}\right)\left(k+\frac{1}{2}\right)\right] \quad (8)$$

with L the frame size, $\tau = t - pL$ and window $h(\tau)$ defined for $\tau = 0, \ldots, 2L-1$. Again, there are a number of fast implementations of the MDCT based on the FFT. The MDCT is one key to success of the ubiquitous "MP3" (MPEG-1 layer III) coding standard, and is now used in the majority of state-of-the-art coding standards, such as MPEG 2/4 AAC.

Using the simple MDCT as described above has severe limitations. Firstly, it is not shift-invariant: at very-low bitrates, this can lead to so-called "warbling artefacts", or "birdies" (as these distortions appear most notably at the higher end of the spectrum). Seondly, the time resolution is limited: for a typical frame size of $L = 1024$ samples at a 44.1 kHz sampling frequency, the resolution is 43 Hz and time resolution is 23 ms. For some very transient signals, such as drums or attacks at note onsets, this value is quite large: this leads to what are known as pre-echo artefacts where the quantization noise "leaks" within the whole window, before the actual burst of energy.

However, the MDCT offers an extra degree of freedom in the choice of the window. This leads to the field of *adaptive (orthogonal) transforms*: when the encoding algorithm detects that the signal is transient in nature, it switches to a "small window" type, whose size is typically 1/8-th of the long window. The transition from long windows to short windows (and vice-versa) is performed by asymmetric windows.

B. Coding in overcomplete bases

Using overcomplete bases for coding may at first seem counter-intuitive, as the number of analysis coefficients is increased. However, we can take advantage of the extra degrees of freedom to increase the sparsity of the set of coefficients: the larger the dictionary, the sparser a solution can be expected. Only those coefficients which are deemed to be significant will be transmitted and coded, i.e. $x(t) \simeq \sum_{\gamma\in\Gamma} u_\gamma \phi_\gamma(t)$, where Γ is a small subset of indices. However, the size of the dictionary cannot be increased at will to increase sparsity, for two reasons. Firstly, solving the inverse problem is computationally intensive and very large dictionaries may lead to overly long computations. Secondly, not only must the values $\{u_\gamma\}_{\gamma\in\Gamma}$ be transmitted, but also the subset $\{\gamma \mid \gamma \in \Gamma\}$ of significant parameters must itself be specified.

In [20], the simultaneous use of $M = 8$ MDCT bases was proposed and evaluated, where the scales (frame sizes) L_m go as powers of two $L_m = L_0 2^m$, $m = 1 \ldots 8$, with window lengths from 128 to 16384 samples (2.9 ms to 370 ms). The 8-times overcomplete dictionary is now $\mathcal{D}_m = \{\bar{\phi}_{mkp} \mid 0 \le p < P_m, 0 \le k < L_M\}$. To reduce pre-echo, large windows are removed from the dictionary near onsets. Finally, the significant coefficients $\{u_\gamma\}_{\gamma\in\Gamma}$ are quantized and encoded together with their parameters $\{\gamma \mid \gamma \in \Gamma\}$. For the sake of efficiency, the sparse decomposition is performed using the Matching Pursuit algorithm [12].

[1] mptk.irisa.fr

Formal listening tests have shown that this coder (named "8*MDCT") outperforms MPEG-2 AAC at very low bitrates (around 24 kbps) for some simple sounds while being of similar quality for complex, polyphonic sounds. At the highest bitrates (above 64 kbps), where a large number of transform coefficients have to be encoded and transmitted, having to encode the extra scale parameter becomes a heavy penalty, and the overcomplete dictionary performs slightly worse than the (adapted) orthogonal basis, although transparency can still be obtained in both cases.

C. New trends

A further key advantage of using overcomplete representations such as "8*MDCT" is that a large part of the information is carried by the significant scale-frequency-time parameters $\{\gamma = (m, k, p) \mid \gamma \in \Gamma\}$, which provide directly interpretable information about the signal content. This can be useful for instance in audio indexing for data mining: if a large sound database is available in an encoded format, a large quantity of user-intuitive information can be easily inferred from the sparse representation, at a very low computational cost. The "8*MDCT" representation was found to have similar performance to the state-of-the-art in common Music Information Retrieval tasks (e.g. rhythm extraction, chord analysis, and genre classification) while MP3 and AAC codecs only performed well in the rhythm extraction, due to poor frequency resolution of those transforms for the other tasks [21].

Sparse overcomplete representations also offer a step towards the "Holy Grail" of audio coding: object coding [22]. In this paradigm, any recording would be decomposed into a number of elementary constituents such as notes, or instruments' melodic lines, that could be rearranged at will without perceivable loss in sound quality. Of course, this if far out of reach for current technology if we make no further assumptions on the signal, as this would imply that we were able to fully solve both the "hard" problems of polyphonic transcription and the underdetermined source separation problem. However, some attempts in very restricted cases [23], [24] indicate that this may be the right approach towards "musically-intelligent" coding.

D. Application to denoising

Finding an efficient encoding of an audio signal based on sparse representations can also help us with *audio denoising*. Typically, while the desired part of the signal is well represented by the sparse representation, noise is typically poorly represented by the sparse representation. By transforming our signal to its sparse representation, discarding the smaller coefficients, and reconstructing the signal again we have a simple way to suppress a significant part of the signal noise.

Many improvements can be made over this simple model. If this is considered in a Bayesian framework, the task is to estimate the most probable original signal given the corrupted observation. Such a Bayesian framework allows the inclusion of structural priors for musical audio objects that take into account the 'vertical' frequency structure of transients and the 'horizontal' structure of tonals, as well as the variance of the residual noise. Such a structured model can help to reduce the so-called 'birdies' or 'musical noise' that can occur due to switching of time-frequency coefficients. However, calculating the sparse representation is more complex than a straightforward Basis Pursuit method, but Markov chain Monte-Carlo (MCMC) methods have been used for this [25].

III. Source Separation

In many applications, audio recordings are mixtures of underlying audio signals and it is desirable to recover those original signals. For example, in a meeting room we may have several microphones, but each one collects a mixture of several talkers. To automatically transcribe the minutes of a meeting, a first step would be to separate these into one channel per talker. Sparse representations can be of significant help in solving this type of source separation problem.

Let us first consider the *instantaneous* mixing model, where we ignore time delays and reverberation. Here we have J audio sources $s_j(t)$, $j = 1, \ldots, J$ which are instantaneously mixed to give I observations $x_i(t)$, $i = 1, \ldots, I$ according to

$$x_i(t) = \sum_{j=1}^{J} a_{ij} s_j(t) + e_i(t) \tag{9}$$

where a_{ij} is the amount of source j that appears in observation i, and $e_i(t)$ is noise on the observation $x_i(t)$. This type of mixture might occur in, for example, pan-potted stereo, where early stereo recordings were produced by changing the amount of each source mixed to the left and right channels without any time delays or other effects[2]. We can also write (9) in vector or matrix notation as

$$\mathbf{x}(t) = \sum_{j} \mathbf{a}_j s_j(t) + \mathbf{e}(t) \qquad \text{or} \qquad \mathbf{X} = \mathbf{AS} + \mathbf{E} \tag{10}$$

where e.g. the matrix $\mathbf{X}$ is an $I \times T$ matrix with columns $\mathbf{x}(t)$ and rows $\bar{x}_i$, and $\mathbf{a}_j$ is the jth column of the mixing matrix $\mathbf{A} = [a_{ij}]$.

If the noise $\mathbf{E}$ is small, the mixing matrix $\mathbf{A}$ is known, and $\mathbf{A}$ is square ($I = J$) and full rank, then we can estimate the sources using $\hat{\mathbf{s}}(t) = \mathbf{A}^{-1}\mathbf{x}(t)$; if we have more observations than sources ($I > J$) we can use the pseudo-inverse $\hat{\mathbf{s}}(t) = \mathbf{A}^{\dagger}\mathbf{x}(t)$. If $\mathbf{A}$ is not known (*blind* source separation) then we could use a technique such as *independent component analysis* (ICA) to estimate it [26].

However, if we have fewer observations than sources ($I < J$), then we cannot use matrix inversion (or pseudo-inversion) to unmix the sources. In this case, called underdetermined source separation [27], [28], we can use sparse representations both to help separate the sources, and, for blind source separation, to estimate the mixing matrix $\mathbf{A}$.

A. Underdetermined separation by binary masking

If we transform the signal mixtures $\bar{x}_i = (x_i(t))_{1 \le t \le T}$ into a domain where they have a sparse representation, it is likely that most coefficients of the transformed mixture correspond to

[2] A more accurate model for acoustic recordings is the convolutive model considered below in Eq. (16)

either none or only one of the original sources. By identifying and matching up the sources present in each coefficient, we can recover the original, unmixed sources. Suppose that our J source signals $\bar{s}_j$ all have sparse representations using atoms $\bar{\phi}_q$ from a full rank $Q \times T$ basis matrix $\mathbf{\Phi}$ (with $Q = T$), i.e.,

$$\bar{s}_j = \sum_q z_{jq} \bar{\phi}_q \qquad 1 \leq j \leq J \tag{11}$$

where z_{jq} are the sparse representation coeffients. In matrix notation we can write $\mathbf{S} = \mathbf{Z}\mathbf{\Phi}$ and $\mathbf{Z} = \mathbf{S}\mathbf{\Phi}^{-1}$.

Now denoting $\mathbf{U} = \mathbf{X}\mathbf{\Phi}^{-1}$ the representation of $\mathbf{X}$ in the basis $\mathbf{\Phi}$, for noiseless instantaneous mixing we have

$$\mathbf{U} = \mathbf{A}\mathbf{Z}. \tag{12}$$

For a simple special case, suppose that $\mathbf{Z}$ is so sparse that *at most one* source coefficient j_q is active at each transform index q, i.e. $z_{jq} = 0$ for $j \neq j_q$. In other words, each column of $\mathbf{Z}$ contains at most one nonzero entry, and the source transformed representations are said to have *disjoint supports*. Then (12) becomes

$$\mathbf{u}_q = \mathbf{a}_{j_q} z_{j_q,q} \qquad 1 \leq q \leq Q \tag{13}$$

so that each vector $\mathbf{u}_q$ is a scaled version of one of the mixing matrix columns $\mathbf{a}_j$. Therefore, when $\mathbf{A}$ is known, for each q we can estimate j_q by finding the mixing matrix column $\mathbf{a}_j$ which is most correlated with $\mathbf{u}_q$:

$$\hat{j_q} = \arg\max_j \frac{|\mathbf{a}_j^{\mathrm{T}} \mathbf{u}_q|}{\|\mathbf{a}_j\|_2} \qquad 1 \leq q \leq Q \tag{14}$$

and we construct a *mask* $\varepsilon_{jq} = 1$ if $j = \hat{j_q}$, 0 otherwise. Therefore using this mask to identify the active sources, and multiplying (13) by $\mathbf{a}_{j_q}^{\mathrm{T}}$ and rearranging we get

$$\hat{z}_{jq} = \varepsilon_{jq} \frac{\mathbf{a}_j^{\mathrm{T}} \mathbf{u}_q}{\|\mathbf{a}_j\|_2} \tag{15}$$

from which we can estimate the sources as $\hat{\mathbf{S}} = \hat{\mathbf{Z}}\mathbf{\Phi}$. Due to the binary nature of ε_{jq} this approach is known as *binary masking*.

Even though the assumption that the sources have disjoint supports in the transformed domain is not satisfied for most real audio signals and standard transforms, the binary masking approach remains relevant to obtain accurate (although non exact) estimates of the sources as soon as they have *almost disjoint supports*, i.e., at each transform index q at most one source j has a non negligible coefficient z_{jq}.

The assumption that the sources have essentially disjoint supports in the transformed domain is highly dependent on the chosen transform matrix $\mathbf{\Phi}$. This is illustrated in Fig. 2 where on top we displayed the coefficients $\bar{z}_j$ of three musical sources (i.e. $J = 3$) in some domain $\mathbf{\Phi}$, below we displayed the coefficients $\bar{u}_i \in \mathbb{R}^2$ of a stereophonic mixture of the sources (i.e., $I = 2$) in the same domain, and at the bottom we displayed the scatter plot of $\mathbf{u}_q$, that is to say the collection of $\{\mathbf{u}_q, 1 \leq q \leq Q\}$.

On the left (Fig. 2-(a)), the three musical sources are playing one after another, and the transform is simply the identity matrix $\mathbf{\Phi} = \mathbf{I}$, which is associated with the so-called Dirac basis. At each time instant t, a single source is active, hence the scatter plot of $\mathbf{u}_q$ clearly displays "spokes", with directions given by the columns $\mathbf{a}_j$ of $\mathbf{A}$. In this simple case, the sources can be separated by simply segmenting their time-domain representation using (14) to determine which source is active at each time instant.

In the middle (Fig. 2-(b)), the three musical sources are playing together, and the transform is still the Dirac basis $\mathbf{\Phi} = \mathbf{I}$. The disjoint support assumption is clearly violated in the time domain, and the scatter plot no longer reveals the directions of the columns $\mathbf{a}_j$ of $\mathbf{A}$. On the right (Fig. 2-(c)), the same three musical sources as in Fig. 2-(b) are displayed but in the time-frequency domain rather than the time domain, using the MDCT transform, i.e., the atoms $\bar{\phi}_q$ are given by (8). On the top we observe that, for each source, many transform coefficients are small while only a few of them are non negligible and appear as spikes. A detailed study would show that these spikes appear at different transform indices q for different sources, so for each transform index there is at most one source coefficient j which is non negligible. This is confirmed by the scatter plot at the bottom, where we can see that the vectors $\mathbf{u}_q$ are concentrated along "spokes" in the directions of the columns $\mathbf{a}_j$ of $\mathbf{A}$.

As well as allowing separation for known $\mathbf{A}$, the scatter plot at the bottom of Fig. 2-(c) also illustrates that sparse representations also allow us to estimate $\mathbf{A}$ from the data, in the blind source separation case. If at most one source coefficient is active at each transform index q, then the directions of the "spokes" in Fig. 2-(c) correspond to the columns of $\mathbf{A}$. Therefore estimation of the columns $\mathbf{a}_j$ of $\mathbf{A}$, up to a scaling ambiguity, becomes a clustering problem which can be addressed using e.g. K-means or weighted variants [27], [29], [30], [31].

Finally, we mention that binary masking can also be used when only one channel is available, provided that at most one source is significantly active at each time-frequency index. However in the single channel case we no longer have a direction $\mathbf{a}_j$ to allow us to determine which source is active on which transform index q. Additional statistical information must be exploited to identify the active sources and build the separating masks $\varepsilon_{jq} \in \{0, 1\}$. For example, non-negative matrix factorization (NMF) or Gaussian Mixture Models (GMMs) of short time Fourier spectra can be used to build non-binary versions of these masks $0 \leq \varepsilon_{jq} \leq 1$, associated with time-varying Wiener filtering [32], [33], [34].

B. Time-frequency masking of anechoic mixtures

Binary masking can also be extended when there is noise, and when the mixture process is convolutive, rather than instantaneous. The convolutive mixing model, which accounts for the sound relections on the walls of a meeting room and the overall reverberation, is as follows:

$$x_i(t) \approx \sum_{j=1}^{J} \sum_{n=-\infty}^{+\infty} a_{ij}(n) s_j(t-n), \qquad 1 \leq i \leq I, \tag{16}$$

where $a_{ij}(n)$ is the mixing filter applied to source j to get its contribution to observation i. In matrix notation we can

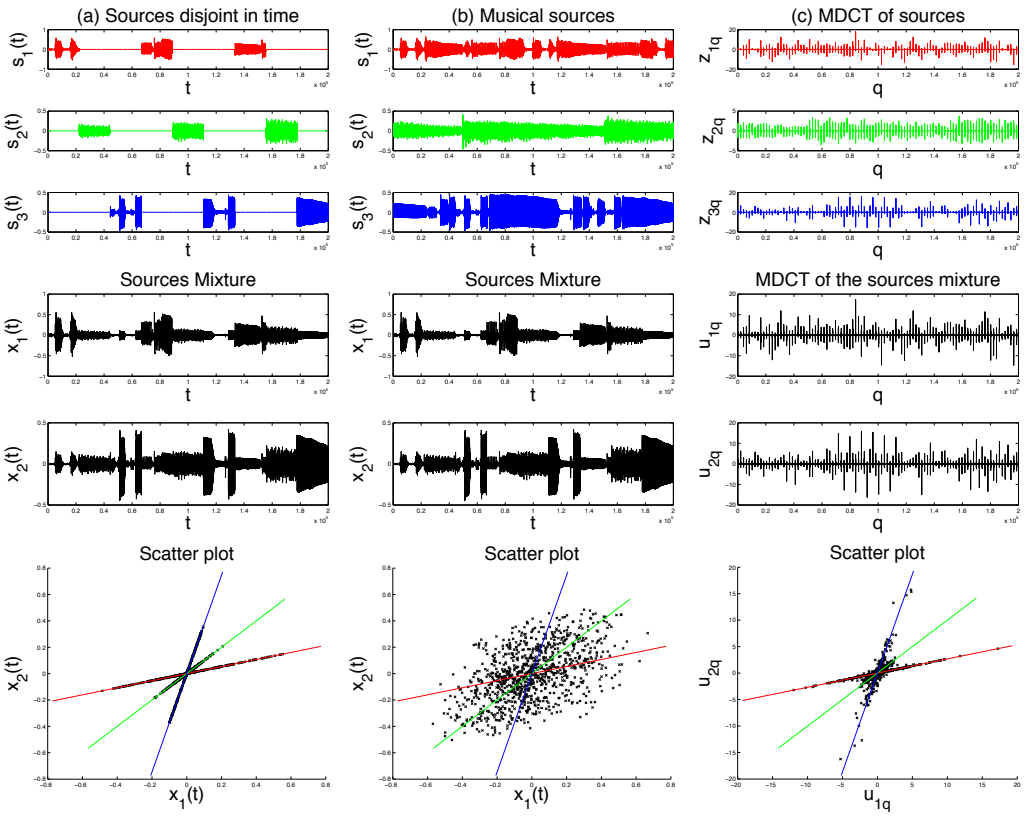

Fig. 2. Top: coefficients of three musical sources. Middle: coefficients of two mixtures of the three sources. Bottom: scatter plot of the mixture coefficients (plain lines indicate the directions $\mathbf{a}_j$ of the columns of the mixing matrix, the colors indicate to which source is associated which column). Left (a): the three musical sources do not play together; time domain coefficients. Middle (b): the three musical sources play together; time domain coefficients. Right (c): the three musical sources play together; time-frequency (MDCT) domain coefficients.

write $\mathbf{X} \approx \mathbf{A} \star \mathbf{S}$, where $\star$ denotes convolution. The STFT of both sides yields an approximate time-frequency domain mixing model [27]

$$X_i(\omega,\tau) \approx \sum_{j=1}^{J} A_{ij}(\omega) S_j(\omega,\tau), \qquad 1 \le i \le I. \tag{17}$$

at frequency ω and time frame τ. For anechoic mixtures, we ignore reverberation but allow different propagation times and attenuations between each source and each microphone. Here the mixing filters $a_{ij}(n)$ become simple gains a_{ij} and delays n_{ij}, giving $A_{ij}(\omega) = a_{ij} \exp(2j\pi n_{ij}\omega)$.

At time-frequency index (ω,τ), suppose that we know that the only significant source coefficients are indexed by $j \in \mathcal{J} = \mathcal{J}(\omega,\tau)$, i.e., $S_j(\omega,\tau) \approx 0$ for $j \notin \mathcal{J}$. Then (17) becomes

$$X_i(\omega,\tau) \approx \sum_{j\in\mathcal{J}} A_{ij}(\omega) S_j(\omega,\tau), \qquad 1 \le i \le I, \tag{18}$$

so that the vectors $\mathbf{u} = \mathbf{u}(\omega,\tau) := (X_i(\omega,\tau))_{i=1}^{I}$ and $\mathbf{z} = \mathbf{z}(\omega,\tau) := (S_j(\omega,\tau))_{j=1}^{J}$ satisfy

$$\mathbf{u} \approx \mathbf{A}_{\mathcal{J}}(\omega)\mathbf{z}_{\mathcal{J}} \tag{19}$$

where $\mathbf{A}_{\mathcal{J}}(\omega) = (A_{ij}(\omega))_{1\le i\le I, j\in\mathcal{J}}$ and $\mathbf{z}_{\mathcal{J}} = (z_j)_{j\in\mathcal{J}}$.

Therefore, for each time-frequency index (ω,τ), if we know the matrix $\mathbf{A}(\omega)$ and the set $\mathcal{J} = \mathcal{J}(\omega,\tau)$ of most significantly active sources, we can estimate the source coefficients as [35]

$$[\hat{S}_j(\omega,\tau)]_{j\in\mathcal{J}} := \mathbf{A}_{\mathcal{J}}^{\dagger}(\omega)\mathbf{u}(\omega,\tau) \tag{20}$$

$$[\hat{S}_j(\omega,\tau)]_{j\notin\mathcal{J}} := 0 \tag{21}$$

where $\mathbf{A}_{\mathcal{J}}(\omega)$ is the mixing filter submatrix for the active sources at frequency ω. Each source can finally be reconstructed by inverse STFT, using e.g. the overlap-add method.

In practice, if we only know the matrix $\mathbf{A}(\omega)$, the critical difficulty is to identify the set $\mathcal{J}$ of most significantly active sources. For a "reasonably small" number of sources with "sufficiently sparse" time-frequency representations, straightforward statistical considerations show that, at most time-frequency points (ω,τ), the total number of active sources is small and does not exceed some $J' \le I$. Identifying the set $\mathcal{J}$ of active source amounts to searching for an approximation $\mathbf{u} \approx \mathbf{A}(\omega)\mathbf{z}$ where $\mathbf{z}$ has few nonzero entries. This is a sparse approximation problem, which needs to be addressed independently at each time-frequency point.

While binary masking corresponds to searching for $\mathbf{z}$ with at most one nonzero entry ($J' = 1$) [27], non-binary masking can be performed choosing, e.g., the minimum 1-norm $\mathbf{z}$ such that $\mathbf{u} = \mathbf{A}(\omega)\mathbf{z}$ (Basis Pursuit) (7), as proposed in [28], or the minimum p-norm solution with $p < 1$ [36].

We have seen in this section that sparse representations can be particularly useful when tackling source separation problems. As well as the approaches we have touched on here

there are many other interesting methods, such as convolutive blind source separation and sparse filter models, which involve sparse representations in the time and/or time-frequency domains. For surveys of some these methods see e.g. [37], [38].

IV. Automatic Music Transcription

So far the coefficients in the sparse representation have been fairly arbitrary, so we were only interested in whether such a sparse representation exists, not specifically what the coefficients mean. However, in some cases, we can assign a specific meaning to the sparse coefficients themselves. For example, in a piece of keyboard music, such as a harpsichord or piano solo, only a few of the many possible notes are playing at any one time. Therefore the notes form a sparse representation when compared to, for example, a time-frequency spectrogram.

In the simplest case, suppose that $\mathbf{x}(\tau) = (X(1,\tau),\ldots,X(\omega,\tau),\ldots,X(K,\tau))^{\mathrm{T}}$ is the spectrum at frame τ. Then we approximate this by the model

$$\mathbf{x}(\tau) \approx \mathbf{A}\mathbf{s}(\tau) = \sum_q \mathbf{a}_q S_q(\tau) \tag{22}$$

where $\mathbf{a}_q$ is the contribution of the spectrum due to note q, and $\mathbf{s}(\tau) = (S_1(\tau),\ldots,S_Q(\tau))^{\mathrm{T}}$ is the vector of note activities $S_q(\tau)$ at frame τ. In this simple case, we are assuming that each note q produces just a scaled version of the note spectra $\mathbf{a}_q$ at each frame τ.

Joining all these spectral vectors together across frames, in matrix notation we get

$$\mathbf{X} \approx \mathbf{A}\mathbf{S}. \tag{23}$$

The basis dictionary $\mathbf{A}$ is no longer of a fixed MDCT or FFT form, but instead must be *learned* from the data $\mathbf{X} = [\mathbf{x}(\tau)]$. To do this, we can use methods such as gradient descent in a probabilistic framework [39], [40] or the recent K-SVD algorithm [41]. When applied to MIDI-synthesized harpsichord music, this simple model is able to identify most of the notes present in the piece, and produce a sparse 'piano-roll' representation of the music, a simple version of *automatic music transcription* (Fig. 3). For more complex sounds, such as those produced by a real piano, the simple assumption of scaled spectra per note no longer holds, and several sparse coefficients are typically needed to represent each note [42].

It is also possible to apply this sparse representations model directly in the time domain, by searching for shift-invariant sparse coding of the musical audio waveforms. Here a 'spiking' representation of the signal is found, which combines with the shift-invariant dictionary to generate the audio signal. For more details and a comparison of these methods, see [43].

V. Conclusions

In this article we have given an overview of a number of current and emerging applications of sparse representations to audio signals. In paritucular, we have seen how we can use sparse representations in audio coding, denoising, source separation, and automatic music transcription. We believe that is an exciting area of research, and we anticipate that there will be many further advances in this area in the future.

Acknowledgements

The authors would like to thank Emmanuel Ravelli for Fig. 1, Simon Arberet for Fig. 2 and Samer Abdallah for Fig. 3.

References

[1] H. Malvar, "A modulated complex lapped transform and its applications to audioprocessing," in *Proc. Int. Conf. Acoust., Speech, and Signal Process. (ICASSP'99)*, vol. 3, 1999.

[2] L. Daudet and B. Torrésani, "Hybrid representations for audiophonic signal encoding," *Signal Processing, special issue on Image and Video Coding Beyond Standards*, vol. 82, no. 11, pp. 1595–1617, 2002.

[3] F. Meyer, A. Averbuch, and R. Coifman, "Multilayered image representation: Application to image compression," *IEEE Trans. Image Process.*, vol. 11, no. 9, pp. 1072–1080, 2002.

[4] J.-L. Starck, M. Elad, and D. L. Donoho, "Redundant multiscale transforms and their application for Morphological Component Analysis," *Journal of Advances in Imaging and Electron Physics*, vol. 132, pp. 287–348, 2004.

[5] S. S. Chen, D. L. Donoho, and M. A. Saunders, "Atomic decomposition by basis pursuit," *SIAM Journal on Scientific Computing*, vol. 20, pp. 33–61, 1998.

[6] I. Daubechies, M. Defrise, and C. De Mol, "An iterative thresholding algorithm for linear inverse problems with a sparsity constraint," *Communications on Pure and Applied Mathematics*, vol. 57, no. 11, pp. 1413–1457, Aug. 2004.

[7] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani, "Least angle regression," *Annals of Statistics*, vol. 32, no. 2, pp. 407–499, Apr. 2004.

[8] M. Figueiredo, J. Bioucas-Dias, and R. Nowak, "Majorization-minimization algorithms for wavelet-based image restoration," *IEEE Trans. Image Process.*, vol. 16, no. 12, pp. 2980–2991, Dec. 2007.

[9] M. Elad, B. Matalon, and M. Zibulevsky, "Coordinate and subspace optimization methods for linear least squares with non-quadratic regularization," *Appl. Comput. Harm. Anal.*, vol. 23, pp. 346–367, 2007.

[10] W. Yin, S. Osher, D. Goldfarb, and J. Darbon, "Bregman iterative algorithms for l1-minimization with applications to compressed sensing," *SIAM J. Imaging Sciences*, vol. 1, no. 1, pp. 143–168, 2008.

[11] P. L. Combettes and J.-C. Pesquet, "A proximal decomposition method for solving convex variational inverse problems," *Inverse Problems*, vol. 24, article ID 065014 (27pp), 2008.

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

[13] Y. C. Pati, R. Rezaiifar, and P. S. Krishnaprasad, "Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition," in *Conference Record of The Twenty-Seventh Asilomar Conference on Signals, Systems and Computers, Pacific Grove, CA*, 1-3 Nov. 1993, pp. 40–44.

[14] D. Needell and R. Vershynin, "Uniform uncertainty principle and signal recovery via regularized orthogonal matching pursuit," *Foundations of Computational Mathematics*, vol. 9, pp. 317–334, 2009.

[15] D. Needell and J. Tropp, "CoSaMP: Iterative signal recovery from incomplete and inaccurate samples." *Applied Computational Harmonic Analysis*, vol. 26, pp. 301–321, 2009.

[16] T. Blumensath and M. Davies, "Iterative hard thresholding for compressed sensing," *Applied and Computational Harmonic Analysis*, 2009, in Press.

[17] S. Krstulovic and R. Gribonval, "MPTK: Matching Pursuit made tractable," in *Proc. Int. Conf. Acoust. Speech Signal Process. (ICASSP'06)*, vol. 3, Toulouse, France, May 2006, pp. III–496–499.

[18] A. Gersho and R. M. Gray, *Vector quantization and signal compression*. Boston: Kluwer, 1992.

[19] S. Mallat, *A Wavelet Tour of Signal Processing: The Sparse Way*, 3rd ed. Academic Press, 2009.

[20] E. Ravelli, G. Richard, and L. Daudet, "Union of MDCT bases for audio coding," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 8, pp. 1361–1372, Nov. 2008.

[21] ——, "Audio signal representations for indexing in the transform domain," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. to appear, 2009.

[22] X. Amatriain and P. Herrera, "Transmitting audio content as sound objects," in *Proceedings of the AES 22nd International Conference on Virtual, Synthetic and Entertainment Audio*, Espoo, Finland, 2002.

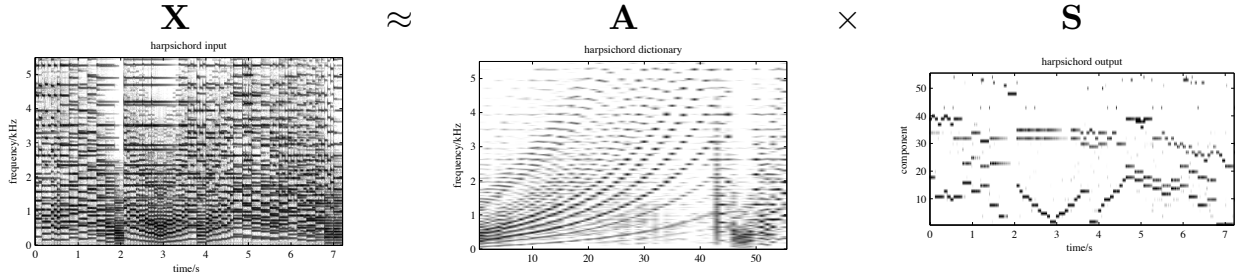

Fig. 3. Transcription of the music spectrogram $\mathbf{X} = [\mathbf{x}(\tau)]$ into the individual note spectra $\mathbf{A} = [\mathbf{a}_q]$ and note activitites $\mathbf{S} = [S_q(\tau)]$ [42].

[23] E. Vincent and M. D. Plumbley, "Low bitrate object coding of musical audio using Bayesian harmonic models," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 4, pp. 1273–1282, May 2007.

[24] G. Cornuz, E. Ravelli, P. Leveau, and L. Daudet, "Object coding of harmonic sounds using sparse and structured representations," in *Proc. of the 10th Int. Conference on Digital Audio Effects (DAFx-07), Bordeaux*, 2007, pp. 41–46.

[25] C. Févotte, B. Torrésani, L. Daudet, and S. J. Godsill, "Sparse linear regression with structured priors and application to denoising of musical audio," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 16, pp. 174–185, 2008.

[26] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent Component Analysis*. John Wiley & Sons, 2001.

[27] O. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Transactions on Signal Processing*, vol. 52, no. 7, pp. 1830–1847, Jul. 2004.

[28] P. Bofill and M. Zibulevsky, "Underdetermined blind source separation using sparse representations," *Signal Processing*, vol. 81, no. 11, pp. 2353–2362, Nov. 2001.

[29] F. Abrard and Y. Deville, "Blind separation of dependent sources using the "TIme-Frequency Ratio Of Mixtures" approach," in *Proc. Seventh International Symposium on Signal Processing and Its Applications (ISSPA 2003)*, vol. 2, Paris, France, Jul. 2003, pp. 81–84.

[30] S. Arberet, R. Gribonval, and F. Bimbot, "A robust method to count and locate audio sources in a stereophonic linear instantaneous mixture," in *Proc. 6th Intl. Conf. on Independent Component Analysis and Blind Signal Separation (ICA 2006), Charleston, SC, USA*, ser. LNCS 3889, J. Rosca *et al.*, Eds. Springer, Mar. 2006, pp. 536–543.

[31] P. Georgiev, F. Theis, and A. Cichocki, "Sparse component analysis and blind source separation of underdetermined mixtures," *IEEE Transactions on Neural Networks*, vol. 16, no. 4, pp. 992–996, 2005.

[32] M. V. S. Shashanka, B. Raj, and P. Smaragdis, "Sparse overcomplete decomposition for single channel speaker separation," in *Proc. Int. Conf. Acoust. Speech Signal Process. (ICASSP'07)*, vol. 2, 2007, pp. II–641–II–644.

[33] L. Benaroya, F. Bimbot, and R. Gribonval, "Audio source separation with a single sensor," *IEEE Trans. Audio, Speech and Language Processing*, vol. 14, pp. 191–199, Jan. 2006.

[34] A. Ozerov, P. Philippe, F. Bimbot, and R. Gribonval, "Adaptation of Bayesian models for single channel source separation and its application to voice / music separation in popular songs," *IEEE Trans. Audio, Speech and Language Processing*, vol. 15, no. 5, pp. 1564–1578, Jul. 2007.

[35] R. Gribonval, "Piecewise linear source separation," in *Proc. SPIE '03*, M. Unser, A. Aldroubi, and A. Laine, Eds., vol. 5207 Wavelets: Applications in Signal and Image Processing X, San Diego, CA, Aug. 2003, pp. 297–310.

[36] E. Vincent, "Complex nonconvex l_p norm minimization for underdetermined source separation," in *Proc. Int. Conf. Indep. Component Anal. and Blind Signal Separation (ICA2001)*. Springer, 2007, pp. 430–437.

[37] 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*, vol. 15, no. 1, pp. 18–33, 2005.

[38] R. Gribonval and S. Lesage, "A survey of sparse component analysis for source separation: Principles, perspectives, and new challenges," in *Proc. 14th European Symposium on Artificial Neural Networks (ESANN'06), 26-28 April 2006, Bruges, Belgium*, 2006, pp. 323–330.

[39] B. A. Olshausen and D. J. Field, "Emergence of simple-cell receptive-field properties by learning a sparse code for natural images," *Nature*, vol. 381, pp. 607–609, 1996.

[40] K. Kreutz-Delgado, J. F. Murray, B. D. Rao, K. Engan, T.-W. Lee, and T. J. Sejnowski, "Dictionary learning algorithms for sparse representation," *Neural Computation*, vol. 15, pp. 349–396, 2003.

[41] M. Aharon, M. Elad, and A. M. Bruckstein, "On the uniqueness of overcomplete dictionaries, and a practical way to retrieve them," *Linear Algebra and its Applications*, vol. 416, pp. 48–67, 2006.

[42] S. A. Abdallah and M. D. Plumbley, "Unsupervised analysis of polyphonic music by sparse coding," *IEEE Trans. Neural Netw.*, vol. 17, pp. 179–196, Jan. 2006.

[43] M. D. Plumbley, S. A. Abdallah, T. Blumensath, and M. E. Davies, "Sparse representations of polyphonic music," *Signal Processing*, vol. 86, no. 3, pp. 417–431, Mar. 2006.

Mark D. Plumbley (S'88–M'90) received the B.A. (Hons.) degree in electrical sciences in 1984 from the University of Cambridge, Cambridge, U.K., and the Ph.D. degree in neural networks in 1991, also from the University of Cambridge. From 1991 to 2001 he was a Lecturer at King's College London. He moved to Queen Mary University of London in 2002, helping to establish the Centre for Digital Music, and where he is now Professor of Machine Learning and Signal Processing and an EPSRC Leadership Fellow. His research focuses on the automatic analysis of music and other audio sounds, including automatic music transcription, beat tracking, and audio source separation, and with interest in the use of techniques such as independent component analysis (ICA) and sparse representations. Prof. Plumbley chairs the ICA Steering Committee, and is an Associate Editor for the IEEE Transactions on Neural Networks.

Thomas Blumensath (S'02–M'06) received the B.Sc. (Hons.) degree in music technology from Derby University, Derby, U.K., in 2002 and the Ph.D. degree in electronic engineering from Queen Mary, University of London, U.K., in 2006. After three years as a Research Fellow in Digital Signal Processing at the University of Edinburgh, he joint the University of Southampton in 2009, where he is currently a Postdoctoral Research Fellow in Applied Mathematics. His research interests include mathematical and statistical methods in signal processing with a focus on sparse signal models and their application.

Laurent Daudet (M'03) was (at the time of writing) Associate Professor at the Pierre-and-Marie-Curie University (UPMC Paris 6), France. After a physics education at the Ecole Normale Superieure in Paris, he received a Ph.D. degree in applied mathematics from the Universit de Provence, Marseille, France, in 2000. In 2001 and 2002, he was a EU Marie Curie Post-doctoral Fellow at the Centre for Digital Music at Queen Mary University of London, UK. Between 2002 and 2009, he has been working at UPMC in the Musical Acoustics Laboratory (LAM), now part of the D'Alembert Institute for mechanical engineering. As of Sept 2009, he is Professor at the Paris Diderot University, with research in the Langevin Institute for Waves and Images (LOA); he is also Visiting Senior Lecturer at Queen Mary University of London, UK. He is author or coauthor of over 70 publications on various aspects of audio digital signal processing, such as audio coding with sparse decompositions.

Rémi Gribonval (M'02–SM'06) graduated from École Normale Supérieure, Paris, France in 1997. He received the Ph. D. degree in applied mathematics from the University of Paris-IX Dauphine, Paris, France, in 1999, and his Habilitation à Diriger des Recherches in applied mathematics from the University of Rennes I, Rennes, France, in 2007. He is a Senior Member of the IEEE. From 1999 until 2001 he was a visiting scholar at the Industrial Mathematics Institute (IMI) in the Department of Mathematics, University of South Carolina, SC. He is now a Senior Research Scientist (Directeur de Recherche) with INRIA (the French National Center for Computer Science and Control) at IRISA, Rennes, France, in the METISS group. His research focuses on sparse approximation, mathematical signal processing and applications to multichannel audio signal processing, with a particular emphasis in blind audio source separation and compressed sensing. Since 2002 he has been the coordinator of several national, bilateral and european research projects, and in 2008 he was elected a member of the steering committee for the international conference ICA on independent component analysis and signal separation.

Mike E. Davies (M'00) received the B.A. degree (honors) in engineering from Cambridge University, Cambridge, U.K., in 1989 and the Ph.D. degree in nonlinear dynamics and signal processing from University College London, London (UCL), U.K., in 1993. He currently holds a SFC funded chair in Signal and Image Processing at the University of Edinburgh, Edinburgh, U.K. His current research interests include: sparse approximation, compressed sensing, and their applications. Prof. Davies was awarded a Royal Society Research Fellowship in 1993 and was an Associate Editor for the IEEE Transactions on Audio, Speech, and Language Processing (2003–2007).