16th European Signal Processing Conference (EUSIPCO 2008), Lausanne, Switzerland, August 25-29, 2008, copyright by EURASIP

COMPRESSED SENSING OF AUDIO SIGNALS USING MULTIPLE SENSORS

Anthony Griffin and Panagiotis Tsakalides

Department of Computer Science, University of Crete and
Institute of Computer Science, Foundation for Research and Technology - Hellas (FORTH-ICS)
Heraklion, Crete, Greece

agriffin@ics.forth.gr and tsakalid@ics.forth.gr

ABSTRACT

Compressed sensing is an attractive compression scheme due to its universality and lack of complexity on the sensor side. In this paper we present a study on compressed sensing of real, non-sparse, audio signals. We investigate the performance of different bases and reconstruction algorithms. We then explore the performance of multi-sensor compressed sensing of audio signals and present a novel scheme to provide improved performance over standard reconstruction algorithms. We then present simulations and measured results of a new algorithm to perform efficient detection and estimation in a sensor network that is used to track the location of a subject wearing a tracking device, which periodically transmits a very sparse audio signal. We show that our algorithm can dramatically reduce the number of transmissions in such a sensor network.

1. INTRODUCTION

Compressed Sensing (CS) [1] [2] seeks to represent a signal using a number of linear, non-adaptive measurements. Usually the number of measurements is much lower than the number of samples needed if the signal is sampled at the Nyquist rate, thus providing the benefits of reduced storage space and transmission bandwidth due to the phenomenal compression achieved.

CS requires that the signal is very *sparse* in some basis—in the sense that it is a linear combination of a small number of the basis functions—in order to correctly reconstruct the original signal. However, the CS measurements made are usually not dependent on the basis used in reconstruction, and thus the measurement process is *universal* as it does not need to change as different types of signals are sensed.

The majority of the literature on CS has been concerned with very sparse signals, and very few results have been presented that explore the performance of CS when used with signals that are not truly sparse. There are even fewer studies on the applicability of CS to audio signals, particulary on speech, music or naturally-occurring signals such as animal calls and environmental sounds. All of these signals are usually not sparse and have a large number of non-zero components in whatever basis might be used in reconstruction.

In this paper we present a study of the performance of CS for a variety of audio signals. We illustrate the differences in performance depending on the basis and the reconstruction algorithm used.

Due to its universality and lack of complexity on the sensor side, CS is an attractive compression scheme for multi-sensor systems. This leads us to investigate the performance of a multi-sensor CS system of real audio signals using standard reconstruction algorithms, and propose a novel scheme to provide improved performance.

We also consider a sensor network that is used to track the location of a subject wearing a tracking device that periodically transmits an audio signal. Here the goal is detection and estimation of the audio signal, which can be done with far fewer measurements than that required for reconstruction. We develop algorithms to perform efficient detection and estimation of these truly-sparse audio signals in a multi-sensor system.

This work was funded by the Marie Curie TOK-DEV "ASPIRE" grant within the 6th European Community Framework Program.

2. COMPRESSED SENSING

2.1 Measurements

We sample the audio signal $x(t)$ at the Nyquist rate and process it in blocks of N samples. Each block is then an $N \times 1$ vector x_k, where k represents the time dependence. The sample vector x_k can be represented as

$$\mathrm{x}_k = \Psi \mathrm{X}_k, \tag{1}$$

where Ψ is an $N \times N$ matrix whose columns are the similarly sampled basis functions $\Psi_i(t)$, and X_k is the vector that chooses the linear combinations of the basis functions. X_k can be thought of as x_k in the domain of Ψ, and it is X_k that is required to be sparse for CS to perform well.

At the sensor, we take M non-adaptive linear measurements of x_k, where $M \ll N$, resulting in the $M \times 1$ vector y_k. This measurement process can be written as

$$\mathrm{y}_k = \Phi_k \mathrm{x}_k, \tag{2}$$

where Φ_k is an $M \times N$ matrix representing the measurement process. For the compressed sensing to work, Φ_k and Ψ must be *incoherent*. In order to provide incoherence that is independent of the basis used for reconstruction, a matrix with elements chosen in some random manner is generally used.

Note that it is not necessary to sample $x(t)$ at the Nyquist rate and then take the discrete measurements. By the use of suitable hardware, it is possible to go straight to the required measurements [3] [4] [5].

2.2 Reconstruction

Once y_k has been measured, it is transmitted in some fashion to a processor, where it is reconstructed. Reconstruction of a compressed sensed signal involves trying to recover the sparse vector X_k. There are two main reconstruction algorithms used: basis pursuit (BP) [1] [6] and orthogonal matching pursuit (OMP) [7] [8]. Let $\hat{\mathrm{X}}_k$ be the recovered signal after reconstruction. BP seeks to find a solution to the following equation

$$\hat{\mathrm{X}}_k = \arg\min \|\mathrm{X}_k\|_1 \qquad \text{s.t.} \qquad \mathrm{y}_k = \Phi_k \Psi \mathrm{X}_k, \tag{3}$$

where $\|\cdot\|_1$ is the ℓ_1 norm. In general, the ℓ_n norm is defined as

$$\|\mathrm{a}\|_n = \left(\sum_j |a_j|^n \right)^{\frac{1}{n}}. \tag{4}$$

Thus BP attempts to find the solution to (3) with the smallest ℓ_1 norm. In doing this, it provides a coefficient for each of the basis functions of Ψ.

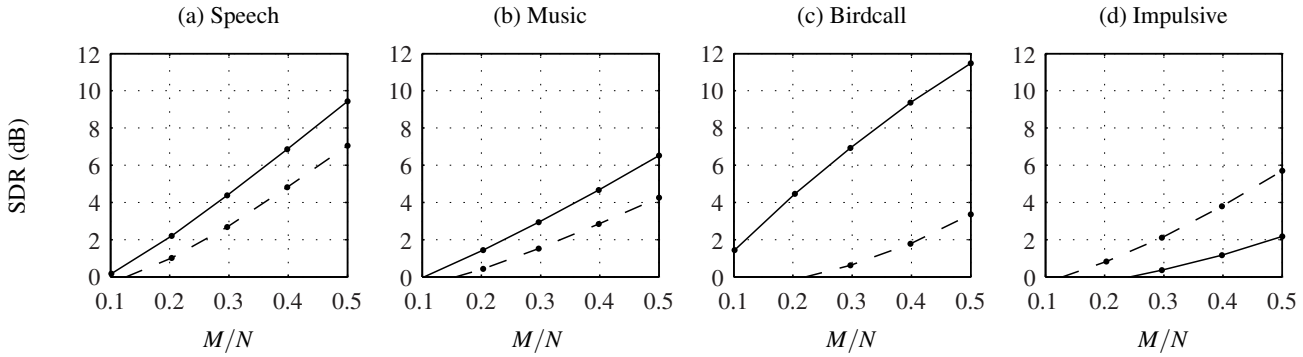

Figure 1: SDR values for reconstructions of a variety of audio signals using basis pursuit and the DCT (solid line) or the DWT (dashed line) for various values of M/N.

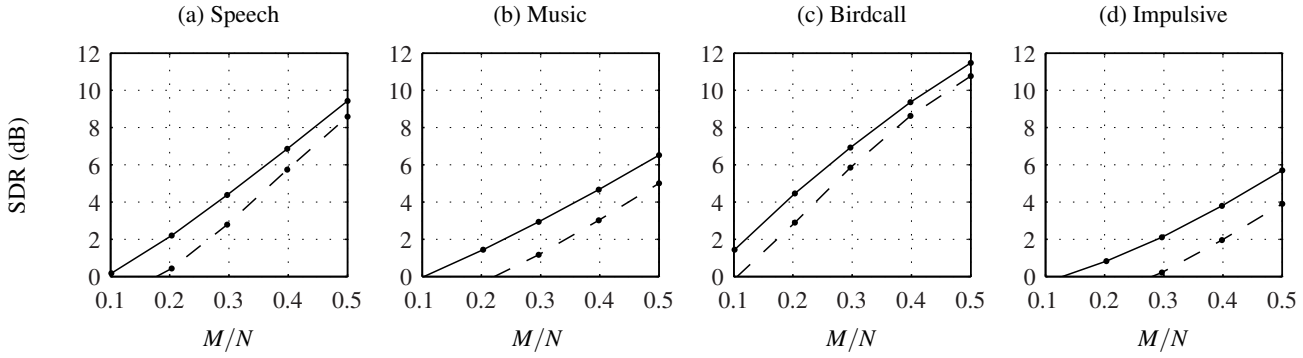

Figure 2: SDR values for reconstructions of a variety of audio signals using basis pursuit (solid line) or orthogonal matching pursuit (dashed line) for various values of M/N.

OMP successively approximates $\hat{\mathrm{X}}_k$ by finding which basis function would contribute the most to

$$\mathrm{y}_k \approx \Phi_k \Psi \hat{\mathrm{X}}_k. \tag{5}$$

It then removes the contribution of this component, orthogonalises the residual, and repeats the process.

Thus BP can be thought of as simultaneously deciding on all of the components of $\hat{\mathrm{X}}_k$, whereas OMP does this on a component-by-component basis. BP also has the advantage of being able to provide a solution that is within a given $||\cdot||_2$ distance, which can be useful for noisy signals .

Nothing comes for free however, and BP is generally more computationally-complex than OMP, but it generally performs better, particularly for signals that are not strictly sparse.

Obviously $\hat{\mathrm{X}}_k$ depends on the choice of the basis Ψ. One of the advantages of CS is that the sensing is the same regardless of the basis used for reconstruction, so that given y_k, $\hat{\mathrm{X}}_{1,k}$ will be recovered if Ψ_1 is used in reconstruction while the use of Ψ_2 would result in $\hat{\mathrm{X}}_{2,k}$. From (1), $\hat{\mathrm{X}}_{1,k}$ and $\hat{\mathrm{X}}_{2,k}$ will give $\hat{\mathrm{x}}_{1,k}$ and $\hat{\mathrm{x}}_{2,k}$ respectively, which could be very different approximations to the original signal x_k.

To illustrate this, Figure 1 presents the signal-to-distortion ratio (SDR) of the recovered signal for various example audio signals for two different bases: the discrete cosine transform (DCT) and a discrete wavelet transform (DWT) with Symmlet filters of order 8. These bases were chosen as they are both orthonormal, and suitable for a variety of audio signals [9]. The signals are the same as those studied in [9], they are sampled at 8kHz, and about 8 seconds in length. We used a block length N of 128, although our experience shows that the SDR of the reconstructed signals is relatively unaffected by the choice of N. The sensing was done using random Gaussian matricies and reconstructed using BP. The SDR is a measure of the degradation due to CS, and is calculated in the following manner.

$$\mathrm{SDR} = \frac{\sum_{k=0}^{K-1} \|\mathrm{x}_k\|_2}{\sum_{k=0}^{K-1} \|\mathrm{x}_k - \hat{\mathrm{x}}_k\|_2}. \tag{6}$$

The M/N on the x-axes of Figure 1 can be thought of as a coding rate, that is, we are representing the full-rate signal with M/N of the number of full-rate samples.

It is clear from Figure 1 that the DCT performs better than the DWT for all but the impulsive signal, which is used as an example of sounds such as sticks breaking and handclaps, sounds that are very time-limited, and may contain many frequencies. The birdcall, in Figure 1(c) is a particularly good example of the importance of using the right basis, as the performance when using the DCT is far superior to that of the DWT.

Figure 2 presents the SDR values of the same signals when reconstructed with BP or OMP. The best performing basis was used for each signal (the DWT for the impulsive signal, and the DCT for the others). With truly sparse signals there is very little difference between the two algorithms, but it is obvious in Figure 2 that BP outperforms OMP for these non-sparse signals. Again, this is due to the fact that BP seeks to estimate the basis components simultaneously, whereas OMP does this in a component-by-component fashion.

3. MULTIPLE SENSORS AUDIO MODEL

Its universality and the fact that very little processing is done on the sensor side and that a greatly reduced number of measurements are required make CS an attractive candidate for use in a sensor network where processing power and transmission bandwidth are usually limited. To this end, we now consider a sensing network of L sensors (microphones) around a sound source.

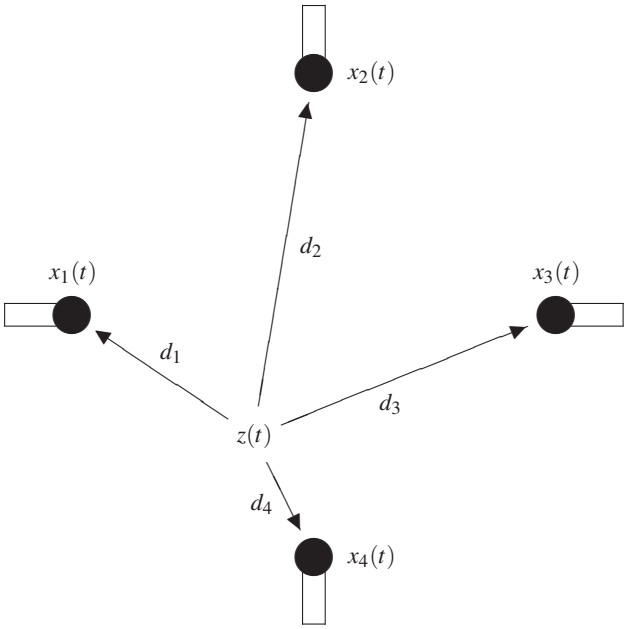

Figure 3: Example set up of an audio four-sensor system, $z(t)$ is the original signal at its source, $x_l(t)$ the signal received at the l-th sensor, and d_l the corresponding path length.

Assuming the sound source is omni-directional and that there are no reflections, the signal received at the l-th sensor will just be a delayed and scaled version of the original signal $z(t)$ at the source

$$x_l(t) = \alpha_l z(t-\tau_l), \tag{7}$$

where α_l and τ_l are the attenuation and delay at the l-th sensor, respectively. Such a model fits the second joint sparsity model (JSM-2) of [10]. Figure 3 shows an example set-up with four microphones. The delay at the l-th sensor is given by $\tau_l = d_l/c$, where d_l is the distance between the source and the l-th sensor, and c is the speed of sound (equal to 344m/s at 21°C). Assuming a point source model, the attenuation at the l-th sensor is given by $\alpha_l = 1/4\pi d_l^2$.

4. CASE STUDY I: RECONSTRUCTION OF REAL, NON-SPARSE AUDIO SIGNALS

4.1 System Architecture

Using the model of Section 3 we investigated the performance of multi-sensor CS of audio signals. We initially used simultaneous orthogonal matching pursuit (SOMP) [10] [11], which is of course based on OMP, and seeks to recover $z(t)$ using M measurements of each of $x_l(t)$, $l = 1,2,\ldots,L$. The idea is to exploit the common structure of the signals at each receiver, and much improved performance can be achieved for very sparse signals.

Unfortunately we found no such improvement for our audio signals, which we surmise is again due to their lack of true sparsity. We found no similar multi-sensor scheme based on BP, so we investigated using a simple scheme to reconstruct the L signals *individually* using BP and then re-combine them. We call this scheme MS-BP.

Let $\hat{\mathrm{x}}_l$ denote the signal from the l-th sensor recovered using BP. Note that there is no time index here as we have concatenated all the blocks of interest into one large vector. We can easily estimate the relative delays between each pair of signals by correlating them over a range of delays, and selecting the delay that maximizes the correlation. Having calculated the relative delays, it is then a simple process of comparisons to determine which sensor is closest to the sound source and therefore has the minimum delay. We will call this signal $\hat{\mathrm{x}}_{l_1}$.

Let $\hat{\mathrm{x}}'_l$ denote a version of $\hat{\mathrm{x}}_l$ that has been time-aligned with $\hat{\mathrm{x}}_{l_1}$ and truncated to the minimum common length of all L signals. We can now calculate the final estimate in MS-BP as

$$\hat{\mathrm{x}} = \frac{1}{L}\left[\hat{\mathrm{x}}'_{l_1} + \sum_{l\neq l_1} \frac{\left\|\hat{\mathrm{x}}'_{l_1}\right\|_2}{\left\|\hat{\mathrm{x}}'_l\right\|_2}\hat{\mathrm{x}}'_l\right], \tag{8}$$

which is essentially the mean of the $\hat{\mathrm{x}}_l$'s once their relative delays and attenuations have been accounted for.

4.2 Simulation Results

In Figure 4 we present some results of MS-BP compared with SOMP for the audio signals we considered in Section 2.2. For comparison purposes, we have also included the BP reconstruction of just the closest sensor. We used a four-sensor system with the sound source off-center, $N = 128$, and with the best basis for each signal. MS-BP only estimates the relative delays to the nearest integer number of samples. Also, a real system would have different noise appearing at each sensor, but as we are investigating ideal performance we did not take noise into consideration.

There is a clear performance improvement in MS-BP over SOMP, which again is due to the fact that we are dealing with real signals that are not truly sparse. In particular, SOMP performs very poorly for the impulsive signal in Figure 4(d). This is because the impulsive signal is best reconstructed with the DWT which is a very time limited basis, and the size of delays commonly experienced with audio signals is hard for the SOMP algorithm to deal with. As MS-BP reconstructs the signals from each sensor individually and time-aligns them *before* recombining them it can perform similarly with signals best represented by DWT or the DCT. Note that the impulsive signal represents an important class of signals, as it would be a signal frequently encountered in a multi-sensor system trying to provide intrusion detection.

There is also an improvement over the reconstruction of just the closest sensor, which is intuitively satisfying, but note that MS-BP and SOMP take L times as many measurements as the single sensor result. Thus although it requires the most computation, MS-BP performs the best, can cover a wider area, has built-in redundancy, and with the reconstruction of the L signals and their relative delays, location detection is also possible.

5. CASE STUDY II: DETECTION AND ESTIMATION OF TRULY-SPARSE AUDIO SIGNALS

5.1 System Architecture

We consider a sensor network that is used to track the location of a subject wearing a tracking device that periodically transmits an audio signal. The sensor network would be arranged in some cell-like structure in an outdoor setting, or else the cells could be rooms of a large building similar to the system considered in [12]. Each cell is modelled as in Section 3. We assume that each sensor has very limited computational capability, and simply transmits its readings to a data fusion centre (DFC). Some initial results of such a system were presented in [13]

As we wish to use CS detection algorithms in the DFC to minimise the number of required transmissions, our audio signal must be very sparse, and our system uses short pulses of a single frequency. This also ensures that the signals appearing at each sensor are jointly sparse, allowing us to use the incoherent detection and estimation algorithm (IDEA) [14] in the DFC to detect the presence of a signal. This allows detection and estimation to be performed with significantly fewer measurements than reconstruction requires.

As IDEA was designed for a single sensor, we had to develop a multi-sensor version of IDEA which we call MS-IDEA. This involved adapting it to use SOMP rather than OMP. This allows for intra-signal compression *without* intra-sensor communication.

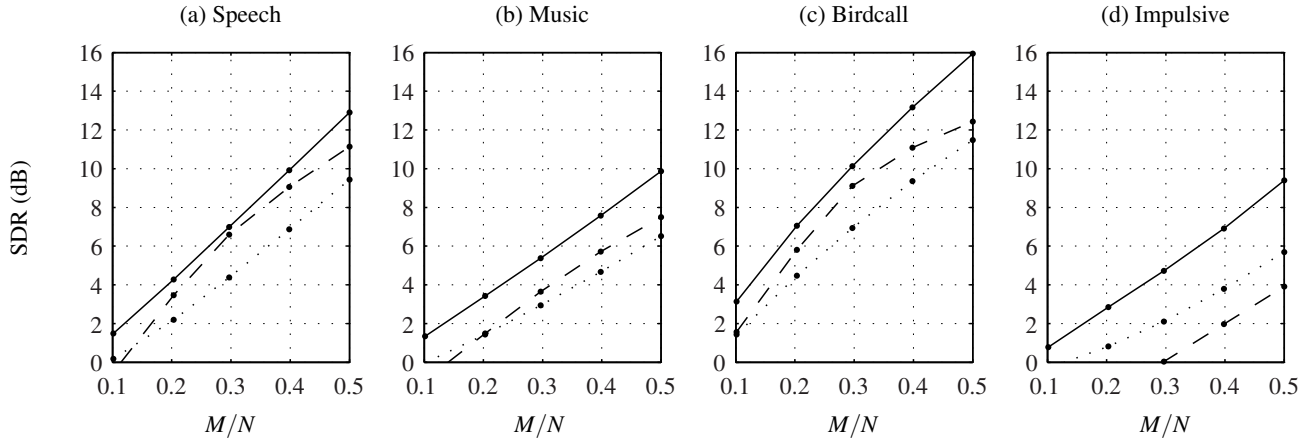

Figure 4: SDR values for reconstructions of a variety of audio signals recorded by a four-sensor system for various values of M/N. The solid line is the MS-BP, the dashed line is SOMP and the dotted line is the BP reconstruction of an individual signal.

Detection in MS-IDEA is performed by testing to see if the highest component after the first iteration of SOMP exceeds a given threshold. Thus the detection process is very computationally efficient, only requiring one iteration of SOMP. In the one component case—as we consider here—estimation in MS-IDEA just involves selecting the index of the component with the highest value after the first pass of SOMP, another very efficient procedure. Note that, as will be shown in the following section, considerably more measurements are required by the estimation process than the detection process to obtain a similar level of performance.

One can thus envisage a scenario where each cell operates in two different modes. A detection mode, where a minimal number of measurements are transmitted to the DFC. Once a subject has been detected in a particular cell, the DFC can instruct the cell to switch to estimation mode, where more measurements are transmitted to the DFC, enabling estimation of the target.

5.2 Simulation Results

From the N samples taken at the Nyquist rate (16kHz) at each sensor node, M are randomly selected and transmitted to the DFC. Thus Φ is formed by randomly selecting M rows of the $N \times N$ identity matrix. Note that with special hardware this could be done in a single process known as Random Sampling [5].

The DFC then uses these LM samples to detect whether or not there is a desired signal in the cell using our MS-IDEA, and if so, then estimate it. The sparsity basis matrix Ψ was an orthonormal inverse FFT matrix.

Figure 5(a) & (b) show the results of a simulation of the detection and estimation performance for a 10×10 metre cell. We simulated the location of the subject on a uniform grid throughout the cell and then calcuated the mean probabilities of detection and estimation over the whole cell.

Curves are given for 1, 2 and 4 sensor nodes at a signal-to-noise ratio (SNR) of 40dB, and it is evident that increasing the number of sensors decreases the number of samples—and therefore transmissions—required per sensor node. For instance, with 4 sensors and a probability of error less than 10^{-2}, only 3 samples are required for detection and 7 for estimation. As $N = 128$, this is equivalent to compressions of 98% and 95%, respectively. Note that this requires extremely little computation on the part of the sensor node.

5.3 Measured Results

Our experimental set-up consisted of two microphones five metres apart and a sound source between them. Pulses of different frequencies were played and recorded at a sampling frequency of 16kHz at each microphone. The sound source was placed at various points between the microphones. The quantities Φ, Ψ and N were the same as in Section 5.2.

We also simulated this configuration, but note that the experiment was performed in a highly-reflective 6×3 metre room and thus there was significant reveberation, which is not taken into account in the simulation. Nevertheless, the results in Figure 5(c)-(f) show that there is good agreement between the simulated and measured results.

The measured results are particularly encouraging as they indicate that MS-IDEA is very suitable for indoor use and that reverberation caused by walls, floors and ceilings does not degrade performance significantly. In fact, the reverberation helps to improve the SNR seen at each sensor; only two samples from both sensor nodes provide a probability of detection error of less than 10^{-2} and one more sample reduces this to less than 10^{-3}.

6. CONCLUSIONS

In this paper we presented a study on the best bases and reconstruction algorithms for compressed sensing (CS) of real, non-sparse, audio signals. We found that basis pursuit reconstruction algorithms out-perform orthogonal matching pursuit algorithms, due to the lack of true sparsity in real audio signals. The choice of basis in reconstruction depends on the signals in question, but this is not of concern for CS due to its universality. We also investigated the performance of a multi-sensor CS system for audio signals, and presented a simple scheme MS-BP to provide improved performance over standard algorithms for a wide range of audio signals, and also provide the possibility of location detection. Through simulations and measurements, we also showed that our algorithm MS-IDEA can be used in a detection and estimation audio sensor network to dramatically reduce the number of transmissions to the DFC. These algorithms require only minimal processing on the sensor side, and only moderate computation in the DFC.

REFERENCES

[1] E. Candès, J. Romberg, and T. Tao, "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information," *IEEE Trans. Inform. Theory*, vol. 52, no. 2, pp. 489–509, February 2006.

[2] D. Donoho, "Compressed sensing," *IEEE Trans. Inform. Theory*, vol. 52, no. 4, pp. 1289–1306, April 2006.

[3] J. Tropp, M. Wakin, M. Duarte, D. Baron, and R. Baraniuk, "Random filters for compressive sampling and reconstruction," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP), Toulouse, France*, vol. 5, May 2006.

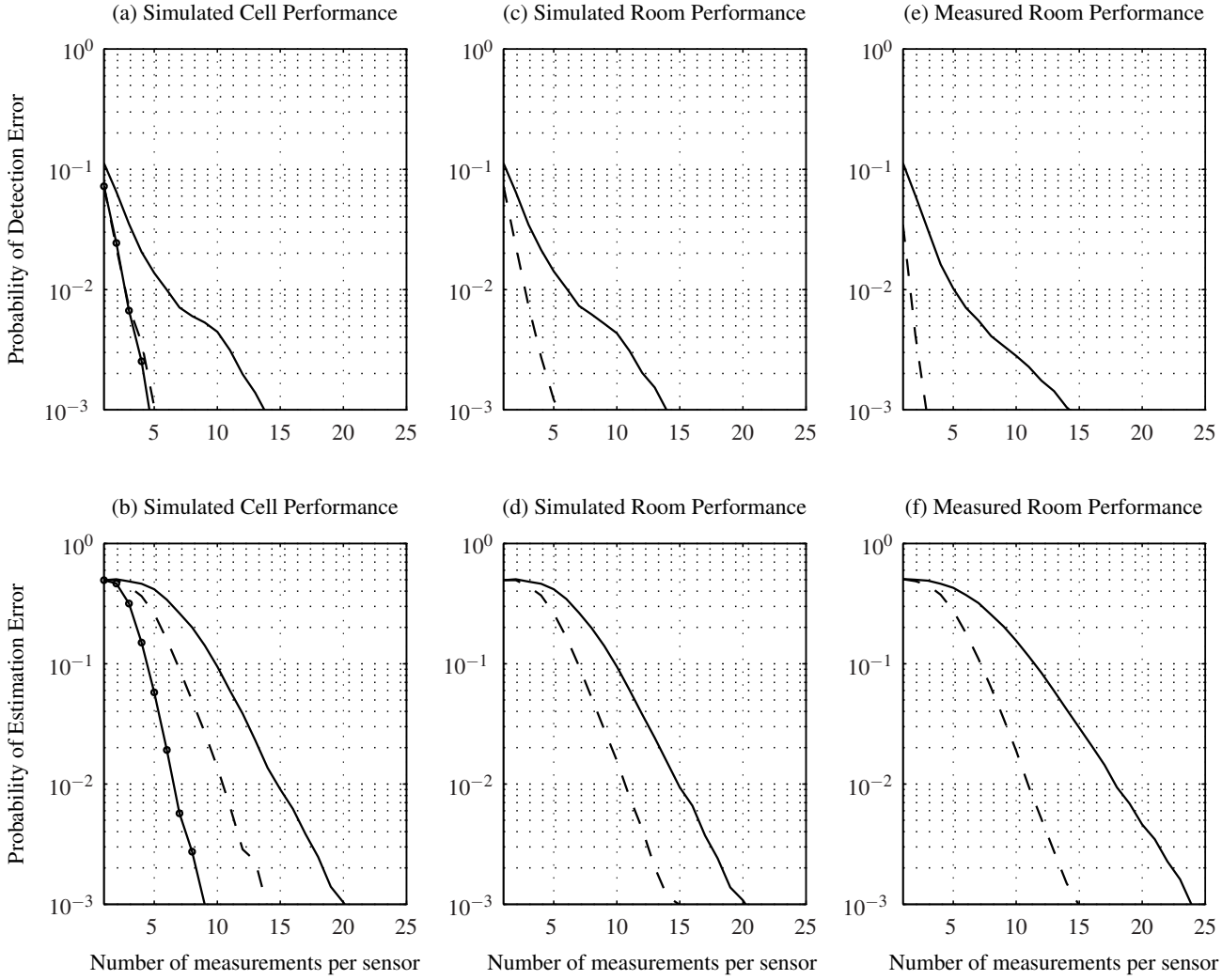

Figure 5: Probability of detection and estimation error for signals recorded by a multi-sensor system. Results for a 10×10 metre cell are shown in (a) & (b), and a 6×3 metre room in (c)-(f). The solid line is the performance using one sensor and IDEA, the dashed line with two sensors and MS-IDEA, and the dotted line with four sensors and MS-IDEA, however this only appears in (a) & (b). The Nyquist rate is 128 samples per sensor.

[4] S. Kirolos, J. Laska, M. Wakin, M. Duarte, D. Baron, T. Ragheb, Y. Massoud, and R. Baraniuk, "Analog-to-information conversion via random demodulation," in *Proc. IEEE Dallas Circuits and Systems Workshop (DCAS), Dallas, TX, USA*, October 2006.

[5] J. Laska, S. Kirolos, Y. Massoud, R. Baraniuk, A. Gilbert, M. Iwen, and M. Strauss, "Random sampling for analog-to-information conversion of wideband signals," in *Proc. IEEE Dallas Circuits and Systems Workshop (DCAS), Dallas, TX, USA*, October 2006.

[6] E. Candès and J. Romberg, "l1-magic: Recovery of sparse signals via convex programming," code package available at www.l1-magic.org.

[7] J. Tropp and A. Gilbert, "Signal recovery from partial information via orthogonal matching pursuit," 2005, preprint.

[8] T. Blumensath, "sparsify," code package available at www.see.ed.ac.uk/∼tblumens/sparsify/sparsify.html.

[9] 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 *Proc. Workshop on Signal Processing with Adaptative Sparse Structured Representations (SPARS'05), Rennes, France*, November 2005.

[10] D. Baron, M. B. Wakin, M. F. Duarte, S. Sarvotham, and R. G. Baraniuk, "Distributed compressed sensing," 2005, preprint.

[11] J. Tropp, A. Gilbert, and M. Strauss, "Simultaneous sparse approximation via greedy pursuit," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP), Philadelphia, PA, USA*, vol. 5, March 2005, pp. 721–724.

[12] P. Y. Chen, W. T. Chen, C. H. Wu, Y.-C. Tseng, and C.-F. Huang, "A group tour guide system with RFIDs and wireless sensor networks," in *Proc. Int. Conf. on Information Processing in Sensor Networks (IPSN), Cambridge, Massachusetts, USA*, April 2007.

[13] A. Griffin and P. Tsakalides, "Compressed sensing of audio signals in a wireless sensor network," in *Proc. European Conference on Wireless Sensor Networks (EWSN), Bologna, Italy*, January 2008.

[14] M. F. Duarte, M. A. Davenport, M. B. Wakin, and R. G. Baraniuk, "Sparse signal detection from incoherent projections," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP), Toulouse, France*, vol. 5, May 2006.