A Perceptually Reweighted Mixed-Norm Method for Sparse Approximation of Audio Signals

Christensen, Mads Græsbøll; Sturm, Bob L.

Published in:
Asilomar Conference on Signals, Systems and Computers. Conference Record

DOI (link to publication from Publisher):
10.1109/ACSSC.2011.6190067

Publication date:
2011

Document Version
Early version, also known as pre-print

A PERCEPTUALLY REWEIGHTED MIXED-NORM METHOD FOR SPARSE APPROXIMATION OF AUDIO SIGNALS

Mads Græsbøll Christensen and Bob L. Sturm

Dept. of Architecture, Design & Media Technology
Aalborg University, Denmark
{mgc,bst}@create.aau.dk

ABSTRACT

In this paper, we consider the problem of finding sparse representations of audio signals for coding purposes. In doing so, it is of utmost importance that when only a subset of the present components of an audio signal are extracted, it is the perceptually most important ones. To this end, we propose a new iterative algorithm based on two principles: 1) a reweighted 1-norm based measure of sparsity; and 2) a reweighted 2-norm based measure of perceptual distortion. Using these measures, the considered problem is posed as a constrained convex optimization problem that can be solved optimally using standard software. A prominent feature of the new method is that it solves a problem that is closely related to the objective of coding, namely rate-distortion optimization. In computer simulations, we demonstrate the properties of the algorithm and its application to real audio signals.

Index Terms— Audio coding, sparse approximations, perceptual distortion measures, audio modeling

1. INTRODUCTION

In recent years, there has been significant interest in methods for finding sparse representations of signals. Such methods aim at decomposing signals into linear combinations of a few vectors from a so-called dictionary. Throughout the past couple of decades, many different methods have been devised for doing this, including matching pursuit [1], basis pursuit [2], the Lasso [3], FOCUSS [4] and many others. The reason for the interest in these methods is that such sparse representations can be used for a wide range of applications, including signal analysis, enhancement, and compression. In audio modeling and coding, which is the topic of the present paper, these ideas extend back to the 80s where so-called parametric models were employed for modeling and modifying musical tones (e.g., [5]). It was, however, in [6] and [7] that these ideas were formalized and cast in the framework of sparse representations. Interestingly, the ideas of using sparse approximations can also be traced back to early speech coders employing linear prediction. Techniques like multi-pulse and regular-pulse excitation use what are essentially sparse models of the prediction error.

In speech and audio applications, it is often the case that we do not wish to reconstruct the observed signal exactly, as would be the case with the representation obtained using basis pursuit. Rather, we wish to represent the signal using a given number of vectors or seek to find the minimal representation that achieves a certain quality. This is, for example, the case in audio coding applications, where restricting the number of vectors would correspond to rate-constrained coding (assuming that the number of bits spent is proportional to the number of vectors used). Similarly, minimizing the number of vectors used to achieve a certain quality corresponds to minimizing the bit-rate required for reconstruction at a certain fidelity in what can be called distortion-constrained coding. In either of these cases, it is of interest that the measure used to quantify the quality of the reconstruction reflects the perceived quality For this reason, efforts have been spent in deriving perceptually motivated norms (e.g., [8]) and in constructing algorithms that use these methods for obtaining sparse representations of audio signals [9–13].

Common to all these methods is that they are based on modifications of matching pursuit-like algorithms to facilitate minimization of an explicit or implicit norm, some being only approximate methods. In this paper, we break this trend by investigating the incorporation of a perceptually motivated norm in methods based on convex optimization by formulating the problem of finding sparse representations of audio signals as constrained optimization problems. More specifically, we adopt and adapt the reweighted 1-norm scheme recently proposed in [14] for our purposes and demonstrate its application to audio modeling and coding. Moreover, we use compressed sensing to obtain an efficient implementation, i.e., as a computational tool rather than as part of the signal acquisition.

The remainder of the paper is organized as follows: In Section 2, the problem of interest is defined mathematically and the proposed method is presented. Then, In Section 3, we present some results obtained using the proposed method before concluding on the work in Section 4.

2. PROPOSED METHOD

2.1. Background

We will now introduce some basic notation and the fundamentals of sparse representations. The problem can be defined as follows. Given a segment of a signal $\mathbf{x} \in \mathbb{R}^N$, in our case and audio signal, and a matrix $\mathbf{Z} \in \mathbb{C}^{N \times F}$ with $F \gg N$ known as the dictionary, we seek to find a sparse coefficient vector $\mathbf{c} \in \mathbb{C}^F$ that recovers $\mathbf{x}$, i.e., $\mathbf{x} = \mathbf{Z}\mathbf{c}$ with a sparse $\mathbf{c}$, or approximately so. To this end, we require two things: 1) a sparsity metric on $\mathbf{c}$ and 2) a reconstruction quality measure. A common way of measuring the sparsity is the 1-norm while the reconstruction quality is most typically measured using the 2-norm. Using these norms, the problem of reconstructing $\mathbf{x}$ with fidelity ψ can be stated as

$$\begin{aligned} &\text{minimize} \, \|\mathbf{c}\|_1 \\ &\text{s.t.} \quad \|\mathbf{x} - \mathbf{Z}\mathbf{c}\|_2 \leq \psi. \end{aligned} \tag{1}$$

In a similar fashion, the problem of minimizing the reconstruction error for a certainty sparsity level can be sated as

$$\begin{aligned} &\text{minimize} \, \|\mathbf{x} - \mathbf{Z}\mathbf{c}\|_2 \\ &\text{s.t.} \quad \|\mathbf{c}\|_1 \leq \psi, \end{aligned} \tag{2}$$

where ψ here is the desired level of sparsity, as measured using the 1-norm. We the solution for any of the considered problems as $\hat{\mathbf{c}}$. There are (at least) two problems with these approaches. Firstly, the 1-norm is not always an accurate measure of sparsity. This can be understood by observing that a vector containing a number of large coefficients is penalized more than a similar vector with small ones. Similarly, a large number of non-zero but small coefficients will not be penalized sufficiently. The second problem is that, as mentioned, the 2-norm is not an accurate measure of the perceived quality of audio reconstructed as $\mathbf{Z}\mathbf{c}$.

2.2. Modifications

To mitigate the problems mentioned above, our solution is twofold: Firstly, we use the re-weighted 1-norm of [14] to measure sparsity. This is based on the principle that if we have an estimate of the coefficient vector, we can form a weighted norm on the coefficient vector in subsequent optimizations such that we have something that is closer to measure sparsity than the 1-norm. Secondly, we use the perceptually weighted 2-norm of [8] as a fidelity measure. The measure of [8] is induced by a perceptual weighting matrix $\mathbf{H} \in \mathbb{R}^{N \times N}$ exhibiting the following circulant structure:

$$\mathbf{H} = \begin{bmatrix} h_1 & h_N & h_{N-1} & \cdots & h_2 \\ h_2 & h_1 & h_N & \cdots & h_3 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ h_{N-1} & h_{N-2} & \vdots & \ddots & h_N \\ h_N & h_{N-1} & h_{N-2} & \cdots & h_1 \end{bmatrix}, \tag{3}$$

which is diagonalized by the Fourier transform and has the reciprocal of the square root of the masking curve as eigenvalues [11]. We note that the resulting matrix is also symmetric, i.e., $h_{N+2-i} = h_i$ for $i > 1$. The perceptual distortion is then measured as $\|\mathbf{H}(\mathbf{x} - \mathbf{Z}\mathbf{c})\|_2$. An interesting and important questions is then what signal to compute $\mathbf{H}$ from. Traditionally, masking curves are computed from the observed signal $\mathbf{x}$, but as argued in [9, 13] it is the masking capabilities of the reconstructed signal $\mathbf{Z}\hat{\mathbf{c}}$ that matters. Consequently, this measure is used in adaptive fashion as in [9, 13]. By computing it from a reconstructed signal in an iterative way, we essentially get a re-weighted 2-norm, and this is what we propose here. It should be noted that this would not be possible with methods like the Lasso since there would be no signal from which to compute $\mathbf{H}$.

The size of the problems generally encountered in audio applications is such that the complexity associated with solving the convex optimization problems is prohibitive. One way to reduce the size of the problems is via compressed sensing with coherent dictionaries [15]. This is done in the following manner. The observed signal $\mathbf{x}$, which contains N samples, is mapped to an M dimensional signal with $M < N$ using random sampling implemented as a so-called measurement matrix $\mathbf{\Phi} \in \mathbb{R}^{M \times N}$. However, since we here apply also a perceptual weighting matrix, which applies to the signal $\mathbf{x}$, this must be done as

$$\mathbf{y} = \mathbf{\Phi}\mathbf{H}\mathbf{x}, \tag{4}$$

i.e., the perceptual weighting matrix must be applied before the random sampling. Similarly, the signal model $\mathbf{Z}\mathbf{c}$ must be modified as $\mathbf{\Phi}\mathbf{H}\mathbf{Z}\mathbf{c}$.

Regarding how to design $\mathbf{\Phi}$ there is a number of issues to consider. Firstly, we are here dealing with an overcomplete dictionary corresponding to the case of coherent dictionaries for which conditions for reconstruction and recovery have been described in [15]. In this case, the measurement matrix $\mathbf{\Phi}$ must satisfy the so-called D-RIP property, which is the case for a Gaussian matrix with M chosen on the order of $C \log(F/C)$ with C being the number of non-zero entries in $\mathbf{c}$. This brings us to the second issue, which is that the measurement matrix must be designed a priori without knowledge of C as the number of partials in the signal varies over time. Here, we adopt the principle of simply using an estimate of C, and, as long as the resulting M is lower than N, this will still result in a reduction of the computation time.

For a general discussion of the applications of compressed sensing to speech and audio signals and the associated challenges and implications, we refer the interested reader to [16].

2.3. The Algorithm

Finally, we will now present the actual algorithm. Let $c_k^{(i)}$ denote the kth element of the vector $\mathbf{c}^{(i)}$ at the ith iteration

and similarly for other quantities. The algorithm performs the following steps for $i = 1, 2, \ldots$:

1. Find the set of coefficients as the solution to the convex optimization problem

$$\begin{aligned} \hat{\mathbf{c}}^{(i)} &= \underset{\mathbf{c}}{\operatorname{argmin}} \|\mathbf{W}^{(i)}\mathbf{c}\|_1 \\ \text{s. t.} \quad & \|\mathbf{\Phi}\mathbf{H}^{(i)}\left(\mathbf{x} - \mathbf{Z}\mathbf{c}\right)\|_2 \leq \psi. \end{aligned} \tag{5}$$

2. Update the weights from $\hat{\mathbf{c}}^{(i)}$ as

$$\mathbf{W}^{(i+1)} = \operatorname{diag}\left(\frac{1}{\left|c_1^{(i)}\right| + \epsilon}, \ldots, \frac{1}{\left|c_F^{(i)}\right| + \epsilon}\right). \tag{6}$$

3. Update the perceptual weighting matrix to obtain $\mathbf{H}^{(i+1)}$ from $\mathbf{Z}\hat{\mathbf{c}}^{(i)}$ using [8].

4. Check convergence. Terminate if converged, otherwise go to step 1.

The algorithm is then terminated when convergence has been achieved, for example when the difference between $\hat{\mathbf{c}}^{(i)}$ and $\hat{\mathbf{c}}^{(i+1)}$ become sufficiently small as measured using a norm. We note that the actual computation of $\mathbf{H}^{(i)}$ requires a rather lengthy description for which reason we do not go into details here but rather simply refer to [8]. The quantity ϵ is a small and positive constant that also can be chosen adaptively [14].

Regarding initialization of $\mathbf{W}^{(1)}$ and $\mathbf{H}^{(1)}$, we initialize $\mathbf{W}^{(1)}$ to an identity matrix while $\mathbf{H}^{(1)}$ is initialized according to the absolute threshold of hearing, corresponding to the masking conditions when no reconstructed signal is available. In relation to the problem stated in (2), we remark that it is straightforward to modify the algorithm accordingly by interchanging the roles of the objective function and the constraints. An interesting aspect of the proposed method is that the constraint ψ is placed on the perceptually weighted error—this means that the constraint reflects a desired quality level rather than the level of the observation noise as is normally the case (see, e.g., [15]). Regarding how to choose ψ, an appealing choice is to use $\psi = \beta\|\mathbf{\Phi}\mathbf{H}^{(i)}\mathbf{x}\|_2$ with $0 < \beta < 1$, meaning that the reconstructed signal should capture at least $1-\beta^2$ of the original signal in terms of perceptually weighted energy, or, equivalently, that the perceptually weighted signal-to-noise ratio should be at least $-20\log_{10}\beta$.

3. RESULTS

We will now report some results illustrating the effectiveness of the proposed method on real signals. The following experiment aims at demonstrating that the proposed method solves a problem that is more meaningful than a direct application of the original reweighted method of [14]. For this purpose, we use an audio signal containing some notes played by a piano. The signal is from the EBU SQAM discs commonly used for assessment of audio coders. The spectrogram of the signal is shown in Figure 1. For visual clarity, we shown only the lower parts of the spectra in the figures. In applying the proposed method, we use a dictionary comprised of windowed complex exponential atoms having uniformly distributed frequencies, segments of 30 ms and a measurement matrix whose entries are realizations of a Gaussian process. Moreover, we have assumed $C = 50$ for each segment and have used $M = 200$. The signal is reconstructed using overlap-add with 50 % overlap, as is commonly done in audio modeling and coding with the signal being windowed similarly to the atoms of the dictionary. The proposed method is implemented in MATLAB using SeDuMi.

The results obtained with the original reweighted method of [14] is shown in Figure 2. Note that his is method is ob-

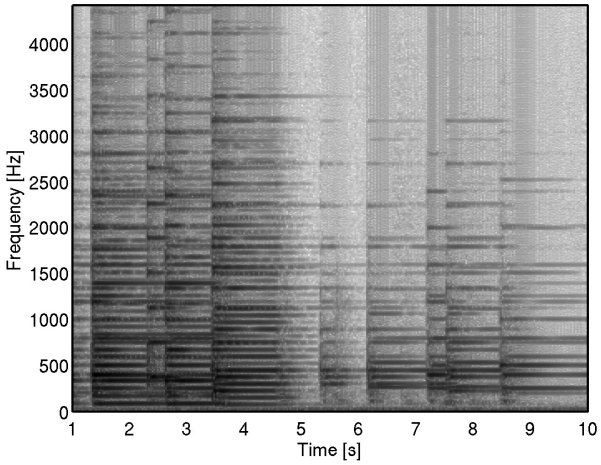

Fig. 1. Spectrogram of the original signal, a piano signal.

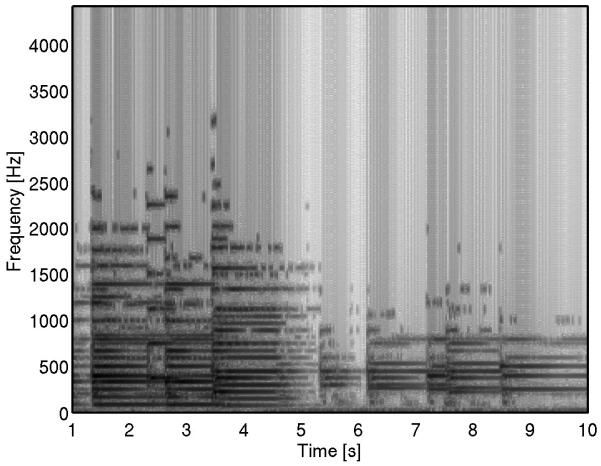

Fig. 2. Spectrogram of the signal in Fig. 1 reconstructed using the reweighted method of [14].

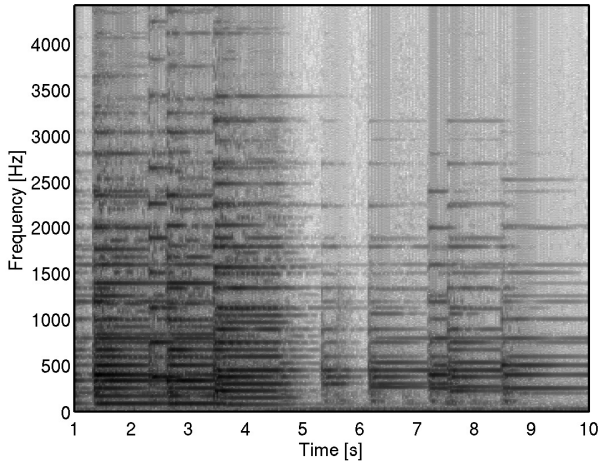

Fig. 3. Spectrogram of the signal in Fig. 1 reconstructed using the proposed method.

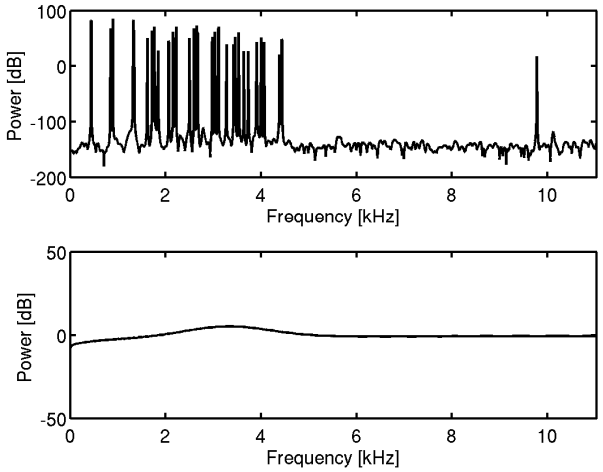

Fig. 4. Entries in the coefficient vector order by frequency (top panel) and perceptual weighting (bottom panel) obtained in iteration 1 using the proposed method applied to a segment of audio.

tained by using $\mathbf{H}^{(i)} = \mathbf{I}, \forall i$ in the algorithm described in Section 2. Both the original and the proposed reweighted methods were run with $\beta = 0.1$ such that they result in the same reconstruction quality, as measured using their respective quality measures. Moreover, at most 10 iterations were used for both methods with the iterations being terminated if $\|\hat{\mathbf{c}}^{(i)} - \hat{\mathbf{c}}^{(i-1)}\|_\infty < 10^{-3}$, as done in [14]. The spectrogram of the signal reconstructed using the proposed method is depicted in Figure 3. Comparing Figures 1, 2, and 3 a number of observations can be made. It can be seen that the perceptual weighting results in more high frequency components being reconstructed well. Moreover, the reconstruction obtained using the proposed can be seen to be much cleaner in the sense that it results in more distinct and more well-separated sinusoidal components in the reconstruction, while the original method tends to cluster components, essentially modeling parts of the signal that are not well-represented using the dictionary (e.g., modulations, onsets, noise), something that leads to artifacts similar to musical noise. These findings were also confirmed by informal listening tests that revealed that the signal reconstructed using the proposed method suffers from much less artifacts than that obtained using the original method. This has also been confirmed to be the case for a large class of signals from the EBU SQAM discs. We note in passing that the comparably poor performance of the method [14] for this signal cannot simply be explained by the use of compressed sensing as the same measurement matrix is used in both cases. Also, it should be noted that in terms of signal-to-noise ratio (measured using the 2-norm), the original reweighted is in fact best achieving the highest score as expected.

To provide some additional insights into the inner workings of the proposed method, the obtained coefficient vector (top panels) order according to the frequency of the atoms along with the perceptual weighting (bottom panels) are shown (both in dB) for a 30 ms segment of an audio signal (in this case a note played by a trumpet) for iterations 1 and 10 in Figures 4 and 5, respectively. The difference, both in terms of the obtained coefficients and the perceptual weighting, is evident. It can be seen that the weighting curve starts out being the reciprocal of the absolute threshold of hearing and the evolves into something more refined where the sinusoidal nature of the signal and the dictionary can be seen. Furthermore, the effectiveness of the reweighted scheme on the sparsity of the coefficients can also clearly be seen as the small coefficients (i.e., the noise floor) are much lower in iteration 10 compared to iteration 1. The effect that the perceptual weighting leads to components being spread more out across frequency can also be seen from these figures. Indeed, it can be seen from Figure 4 that the algorithm at first tends to select clusters of components while this phenomenon cannot be observed in Figure 5.

4. CONCLUSION

In this paper, a new method has been proposed for the purposes of sparse approximation of audio signals. For such signals, it is not only important that a sparse approximation is found but also that it leads to a reconstruction having a high perceived quality. The method employs a 1-norm based measure of sparsity of the coefficient vector and a 2-norm based perceptually motivated measure on the reconstruction fidelity. The method builds on the idea of applying an adaptive weighting of the 1-norm on the coefficient vector in each iteration, where the weighting is chosen as the reciprocal of the vector from the previous iteration. In this manner, the sparsity is enhanced in successive iterations. Similarly, the new algorithm

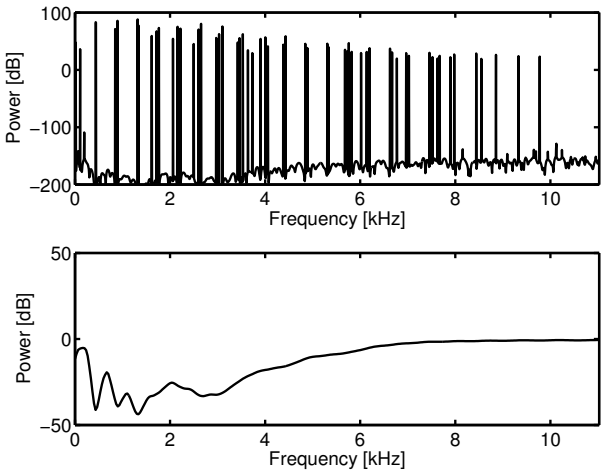

Fig. 5. Entries in the coefficient vector order by frequency (top panel) and perceptual weighting (bottom panel) obtained in iteration 10 using the proposed method applied to a segment of audio.

employs an adaptive weighting in the 2-norm used for measuring the quality of the reconstructed signal. This is done using a perceptual distortion measure such that the masking capabilities of the reconstructed signal is taken into account rather than those of the original signal. This is only possible in the present framework due to the iterative nature of the algorithm. Experiments have confirmed that this leads reconstructions having a higher perceived quality than without the perceptual weighting.

5. REFERENCES

[1] S. Mallat and Z. Zhang, "Matching pursuit with time-frequency dictionaries," *IEEE Trans. Signal Process.*, vol. 41(12), pp. 3397–3415, Dec. 1993.

[2] S. S. Chen, D. L. Donoho, and M. A. Saunders, "Atomic decomposition by basis pursuit," *SIAM J. Sci. Comput.*, vol. 20, pp. 33–61, 1996.

[3] R. Tibshirani, "Regression shrinkage and selection via the lasso," *J. of Royal Stat. Soc.*, vol. 58(1), pp. 267–288, 1996.

[4] I. F. Gorodnitsky and B. D. Rao, "Sparse signal reconstruction from limited data using FOCUSS: A re-weighted minimum norm algorithm," *IEEE Trans. Signal Process.*, vol. 45(3), pp. 600–616, Mar. 1997.

[5] E. B. George and M. J. T. Smith, "Analysis-by-synthesis/overlap-add sinusoidal modeling applied to the analysis-synthesis of musical tones," *J. Audio Eng. Soc.*, vol. 40(6), pp. 497–516, 1992.

[6] M. M. Goodwin, "Adaptive Signal Models: Theory, Algorithms, and Audio Applications," Ph.D. dissertation, University of California, Berkeley, 1997.

[7] R. Gribonval, P. Depalle, X. Rodet, E. Bacry, and S. Mallat, "Sound signal decomposition using a high resolution matching pursuit," in *Proc. Int. Computer Music Conf.*, Aug. 1996, pp. 293–296.

[8] S. van de Par, A. Kohlrausch, R. Heusdens, J. Jensen, and S. H. Jensen, "A perceptual model for sinusoidal audio coding based on spectral integration," *EURASIP J. on Applied Signal Processing*, vol. 9, pp. 1292–1304, 2004.

[9] R. Heusdens and S. van de Par, "Rate-distortion optimal sinusoidal modeling of audio using psychoacoustical matching pursuits," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 2002, pp. 1809–1812.

[10] T. S. Verma and T. H. Y. Meng, "Sinusoidal modeling using frame-based perceptually weighted matching pursuits," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 2, 1999, pp. 981–984.

[11] M. G. Christensen and S. H. Jensen, "On perceptual distortion minimization and nonlinear least-squares frequency estimation," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 14(1), pp. 99–109, Jan. 2006.

[12] M. G. Christensen and S. H. Jensen, "The cyclic matching pursuit and its application to audio modeling and coding," in *Rec. Asilomar Conf. Signals, Systems, and Computers*, 2007, pp. 550–554.

[13] R. Vafin, S. V. Andersen, and W. B. Kleijn, "Exploiting time and frequency masking in consistent sinusoidal analysis-synthesis," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 2, 2000, pp. 901–904.

[14] E. J. Candés, M. B. Wakin, and S. P. Boyd, "Enhancing sparsity by rewighted l_1 minimization," *The Journal of Fourier Analysis and Applications*, vol. 14(5), pp. 877–905, Dec. 2008.

[15] E. J. Candés, Y. C. Eldar, and D. Needell, "Compressed sensing with coherent and redundant dictionaries," *Applied and Computational Harmonic Analysis*, 2010, in press.

[16] M. G. Christensen, J. Østergaard, and S. H. Jensen, "On compressed sensing and its application to speech and audio signals," in *Rec. Asilomar Conf. Signals, Systems, and Computers*, 2009, pp. 356–360.