QUANTIZED SPARSE APPROXIMATION WITH ITERATIVE THRESHOLDING FOR AUDIO CODING

M. Yaghoobi, T. Blumensath and M. Davies

Institute for Digital Communications,
Joint Research Institute for Signal and Image Processing,
University of Edinburgh, UK

ABSTRACT

Sparse coding is a new field in signal processing with possible applications to source coding. In this paper we present a new method that combines the problems of sparse signal approximation with coefficient quantization. This method uses overcomplete dictionaries and exploits signal redundancy. The proposed method will be derived as an extension of a recently presented method (iterative thresholding) to find sparse representations of signals. Because in digital communication and storage we need a quantized representation of the signal, instead of quantization of sparse representations a posteriori, we propose a refined method that combines sparse approximation and quantization. To compare the proposed method to a posteriori quantization, we present an audio example.

Index Terms – Sparse approximation, Quantization, Iterative Thresholding, Audio coding, Signal representation

1. INTRODUCTION

Sparse approximations represent signals with a small number of elementary functions (atoms) from an overcomplete set of functions (dictionary). This kind of signal representation has various applications such as source separation, denoising, feature extraction, compression and source coding. The focus of this paper is to simultaneously obtain a sparse and quantized representation of a signal. As an example, we use an audio signal to show the performance of the algorithm. Sparse representations are potentially useful in source coding because the encoder only needs to encode non-zero coefficients and their indices (i.e. the indices of the atoms in the dictionary) to enable the decoder to reconstruct the original signal.

Most modern audio codecs use a transformation of the input as the first step to get a sparser representation of the signal and, with some psychoacoustic considerations, quantize and encode the coefficients. The decoder uses the inverse transform [1]. The idea behind transform coding is that a simlpe scalar quantizer can be used. Therefore, many researchers use sparse representations based on overcomplete dictionaries to increase the sparsity of the representation (with an increase in the cost of index coding)[2] [3] [4].

Previous approaches mostly use greedy algorithms like Matching Pursuit (MP) or its extension, Quantized MP (QMP) [5], which was shown to improve quantized SNR by 0.5–2 dBs for a fixed bit rate [6]. In this paper we propose a different in-loop quantization method and show that it uses the redundancy in the dictionary to find a better quantized approximation. The contribution of this paper is an iterative algorithm that jointly optimizes the selection of atoms from a redundant dictionary and the quantization. A new penalty function will be presented to replace the traditional penalty function based solely on the number of none-zero coefficients. To optimize this penalty function we need either relaxation or approximation. In this work the latter one is chosen.

2. SPARSE APPROXIMATION AND ITERATIVE THRESHOLDING

An optimal source code can be achieved by Vector Quantization (VQ) [7] which is computationally expensive. Transform coding is used to get suboptimal source codes with simpler algorithms. In standard transform coding, coefficients are quantized with a scaler quantizer and then entropy coded [8]. Linear transforms do not always lead to good performance. One solution is to represent the signals using a nonlinear transform and an overcomplete set of elementary functions. Nonlinear transforms can lead to sparser representations for coding. Overcomplete signal representations can be formulated as,

$$y = Kx, \tag{1}$$

where K is an N by M matrix with $M > N$ and $|K| = N$. y and x are the input signal and the signal in the transform domain. Because K is a non-square matrix with $M > N$, we have an infinite number of solutions x for every input y. We can choose a particular solution based on the constrained

This work is funded by EPSRC grant number D000246.The authors acknowledge their support of the Joint Research Institute with the Heriot-Watt University as a component part of Edinburgh Research Partnership.

optimization of the desired penalty function $P(x)$,

$$\min_{x:y=Kx} P(x) \tag{2}$$

For sparse representations, $P(x)$ is often chosen to be l_0, which measures the number of non-zero coefficients. Instead of solving this exact representation problem, we use an additive cost function of a squared error approximation and the penalty,

$$\min_x \Phi(x) \quad ; \quad \Phi(x) = ||Kx - y||^2 + \lambda P(x) \tag{3}$$

where $|| \cdot ||$ is the norm in signal space and λ a Lagrangian multiplier. In general, solving the above optimization problem based on the l^0 sparsity constraint is an NP-hard problem and is not computable in an acceptable amount of time. So the problem needs to be sinplified using relaxation or approximation [9]

Recently Daubechies et al. [10] have presented an Iterative Thresholding algorithm (IT), as an iterated version of classical thresholding [11] to find sparse approximations for a broader ranges of dictionaries (the classical one was presented for orthogonal wavelets and could be extended to other orthogonal bases). The algorithm was shown to solve a relaxed version of the l^0 problem (with a convex penalty function). The penalty function in [10] is,

$$P(x) = |x|_p^p \tag{4}$$

where $|x|_p$ is the p-norm with $1 \leq p \leq 2$ to ensure convexity of $P(x)$.

The matrix K couples the coefficients and prevents us from optimizing the cost function element-wise. This coupling can be removed by adding a convex function to the cost function, to get a "surrogate function". We can then optimize the new cost function (this process is called optimization transfer),

$$\Phi^S(x, x') = \Phi(x) + ||x - x'||^2 - ||Kx - Kx'||^2 \tag{5}$$

When $x = x'$, the surrogate function is equal to the original cost function. Rewriting (5) yields,

$$\Phi^S(x, x') = \sum_i [(x_i - \alpha_i)^2 + \lambda |x_i|^p] + [\beta - \alpha_i^2]. \tag{6}$$

where $\alpha = (I - K^*K)x' + K^*y$, $\beta = ||y||^2 + ||x'||^2 - ||Kx'||^2$, i shows the element number and K^* is the conjugate transpose of K. α is a function of x also known as a Landweber update of x [12], which could be used iteratively to compute the l^2 regularized optimal solution of the inverse problem. The second term is constant and we only need to optimize the first sum, which is now decoupled and can be minimized elementwise. In an iterative scheme we set the previous computed value, x^{n-1}, to x' and then set x^n to the value x that optimizes

$$\Phi^S(x^n, x^{n-1}, i) = (x_i^n - \alpha_i^{n-1})^2 + \lambda |x_i^n|^p \tag{7}$$

where α^{n-1} is the Landweber update of x^{n-1}. The convergence of this algorithm to a minimum of (3), for certain cost functions, is shown in [10]. In each step we find the best value for x_i^n based on x_i^{n-1} (or its corresponding Landweber update). Therefore the iterative algorithm for M iterations is as follows:

1. $n = 1, x^0 = 0,$
2. $\alpha^{n-1} = (I - K^*K)x^{n-1} + K^*y,$
3. $x_i^n = f(\alpha_i^{n-1}); \forall i$
4. $n = n + 1$ if $n \leq M$ return to step 2.

In step 3, f is the element-wise optimizer. When $p = 1$ and $p = 0$ this function is soft- and hard- thresholding [11], respectively.

The IT algorithm is flexible and it is possible to change the penalty function (albeit under certain conditions). In this paper we propose a Quantized IT algorithm based on certain modifications of the cost function, such that we simultaneously get a quantized signal representation.

3. QUANTIZED SPARSE APPROXIMATION

In this section we are considering the problem of quantized sparse representations. For coding, coefficients need to be quantized. Therefore the transform is changed to get quantized coefficients to reduce quantization error. The quantized version of (3) is:

$$\Phi_Q(z) = ||Kz - y||^2 + P_Q(z) \tag{8}$$

$P_Q(z) = \lambda ||z||_0$ measures the number of non-zero coefficients and z is a quantized value vector with the desired uniform quantizer, with larger zero bin (δ_0 and δ_1 are the zero and non-zero bin sizes). Optimizing the above cost function is an NP-hard problem. But with iterative thresholding in the quantized domain we could decrease this cost function progressively. After adding quantized version of the previously mentioned convex function, the following surrogate function should be minimized in each step:

$$\Phi^S(z^n, z^{n-1}, i) = (z_i^n - \alpha_i^{n-1})^2 + \lambda |z_i|^0 \tag{9}$$

Here $|z_i|^0$ is equal to zero if $z_i = 0$ and equal to one otherwise. We are looking for the optimum value of Φ^S in the quantized value domain. For different z_i^n, Φ^S is

$$\Phi^S(z^n, z^{n-1}, i) = \begin{cases} (\alpha_i^{n-1})^2 & z_i^n = q_0 = 0 \\ (\alpha_i^{n-1} - q_k)^2 + \lambda & z_i^n = q_{k:k \neq 0} \end{cases} \tag{10}$$

where q_k is the k^{th} quantization level ($k \in Z, -\lfloor L/2 \rfloor + 1 \leq k \leq \lfloor L/2 \rfloor$ for a L level quantizer). To define the neighborhood of each q_k in which the optimum value of Φ^S (for the quantized value z_i^n) is q_k , we just need to compare it with Φ^S at adjacent quantization point(s) (q_{k-1} and q_{k+1}). This leads

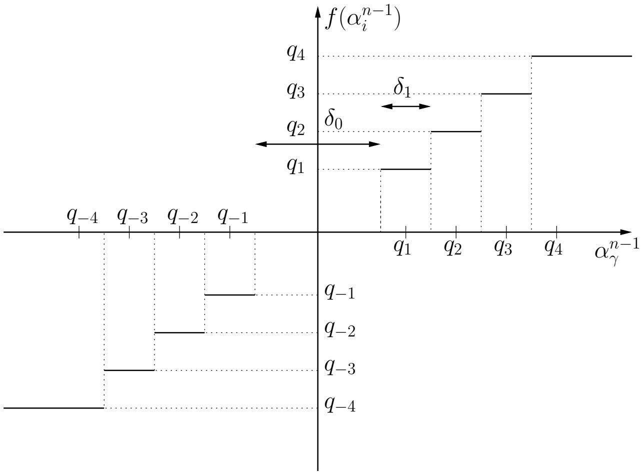

Fig. 1. 9 level on-center QShrinker

to a function on α that is a quantizer with the same quantization levels as the original quantization levels and an adjustable zero bin. We can choose an appropriate λ, by using equation (11), to ensure that the quantizer is uniform in non-zero bins and has a larger zero-bin size, see Figure 1.

$$\lambda = (\delta_0/2)^2 - (\delta_1/2)^2 \tag{11}$$

Therefore the shrinking function changes to a simple uniform quantizer,

$$f(\alpha) = Q(\alpha) \tag{12}$$

With different initial values, the algorithm will converge to different fixed points. Increasing the number of quantization levels directly increase the number of local minima. To improve performance, we adopt a relaxation strategy for iterative shrinkage previously presented in [13]. Instead of updating the current coefficients with the proposed threshold, we choose a relaxation factor μ and update the current coefficients as

$$x_i^n = (1-\mu)x_i^{n-1} + \mu f(\alpha_i^{n-1}), \tag{13}$$

where $0 < \mu \leq 1$. With this update, x_i is not quantized. But it is obvious that the fix points of (12) and (13) are the same. After the algorithm converges, all x_is have quantized values. When $||K|| > 1$, for some initial values, updating by (12) is unstable. But with the use of this relaxation, and choosing appropriate μ, our simulations show stability for both methods (IT and QIT). The overall process is the same as IT but with step 3 replaced by (13).

4. SIMULATIONS

A segment of pop music sampled at 32kHz was chosen here as a test signal (Figure 2). A 4 times overcomplete MDCT dictionary (overcomplete in the frequency domain) was used.

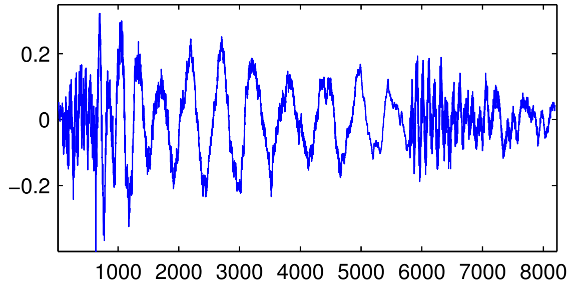

Fig. 2. Input audio signal

All simulations were started with $x^0 = 0$. We fixed quantization levels and used a uniform dead-zone quantizer with the following zero bin to non-zero bin ratio,

$$\zeta = \frac{\delta_0}{\delta_1} \tag{14}$$

By changing ζ, the results of the algorithm will have a varying umber of non-zero coefficients (it should be noted that this convention is not just for QIT. It is also used for IT, where the zero bin is the thresholding parameter. So we can compare equivalent coefficients quantized with QIT for a specific number of non-zero coefficients). A four bit quantizer (16 levels) was selected to quantize each coefficient. To show the convergence of the algorithm, simulations were run for 20 and 100 iterations. The results are shown in Figure 3. The graph with plus symbols is iterative hard thresholding and the results achieved when quantizing this solution are shown with circles. QIT and its quantized output are shown with cross and star symbols. Note that due to the relaxation approach used, the output of QIT is not automatically quantized. The horizontal axis shows the number of non-zero coefficients. We can see that for differenet numbers of non-zero coefficients, IT gives better SNR than QIT. However after quantization of the coefficients, the SNR of the decoded quantized coefficients of QIT is better than quantized IT. We also see that with more iterations, QIT and its quantized output get closer to each other, which shows that the algorithm is converging to a quantized solution. Another observation to be made here is that the SNR starts to decrease when we use a large number of non-zero coefficients. This is an artifact in the analysis where we use a fixed coding cost, i.e. a fix number of quantization levels. To show the benefit of using QIT, we need to show the operating rate-distortion (R-D) curve by computing the convex hull for different bit budgets. The audio sample used in the previous experiment is here used for coding with 4 to 9 bit quantizers. The operational R-D is shown in Figure 4. The graph shows that we have 0.2 dB SNR improvement for 1 bit/sample and up to 1 dB improvement for 12 bits/sample.

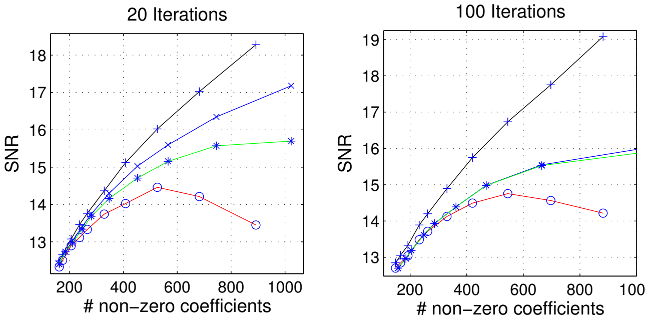

Fig. 3. For two different numbers of iterations (20 and 100) output SNRs are shown in four different cases (IT (+), QIT (x), quantized QIT (*), quantized IT (o))

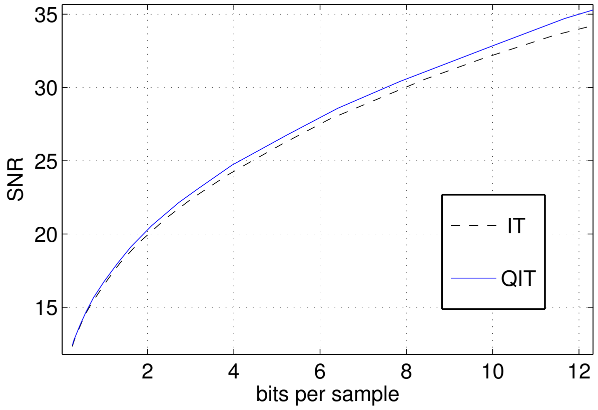

Fig. 4. Operating R-D curves for QIT (upper) and IT (lower)

5. CONCLUSION

In this paper we introduced a new method for jointly approximating and quantizing a signal. The newly presented iterative thresholding method was refined for this purpose and we have shown that even with a small number of iterations the algorithm can give a relatively good result (close to the fixed point). The algorithm is much faster than previously used MP type algorithms. Each iteration of MP and QIT have the same order of computation. However MP extracts one element at a time and therefore requires at least as many iterations as the number of atoms to be extracted, while QIT calculates all the coefficients with less iterations. With the proposed method, we have shown that jointly quantized and sparsified coefficients achieve a better SNR for the same number of non-zero coefficients than sparse approximation and quantization done separately. Because a psychoacoustic model was not considered, this kind of coefficient coding is not comparable with some well known available coders. This paper aims to show the preference of using quantized sparse approximation instead of a posteriori quantization of sparse representation. More investigations are required to study ways of choosing the relaxation parameter, finding an appropriate initialization, considering psychoacoustic models and using listening test for final evaluation.

6. REFERENCES

[1] M. Bosi, *Introduction to Digital Audio Coding and Standards*, Springer, 2002.

[2] M.M. Goodwin, *Adaptive Signal Models: Theory, Algorithms and Audio Applications*, Ph.D. thesis, University of California, Berkeley, 1997.

[3] P. Frossard, P. Vandergheynst, R.M. Figueras i Ventura, and M. Kunt, "A posteriori quantization of progressive matching pursuit streams," *IEEE Trans. on Signal Processing*, vol. 52, no. 2, pp. 525–535, 2004.

[4] M. Davies and L. Daudet, "Sparse audio representations using the MCLT," *Signal Processing*, vol. 86, pp. 457–470, 2006.

[5] V.K. Goyal, M. Vetterli, and N.T. Thao, "Quantized overcomplete expansions in R^N : Analysis, synthesis and algorithms," *IEEE Trans. on Information Theory*, vol. 44, no. 1, pp. 16–31, 1998.

[6] C.D. Vleeschouwer and A. Zakhor, "In-loop atom modulus quantization for matching pursuit and its application to video coding," *IEEE Trans. on Image Processing*, vol. 12, no. 10, pp. 1226–1242, 2003.

[7] A. Gersho and R.M. Gray, *Vector Quantization and Signal Compression*, Kluwer Academic Publishers, 1991.

[8] V.K. Goyal, "Theoretical foundations of transform coding," *IEEE Signal Processing Magazine*, vol. 18, no. 5, pp. 9–21, 2001.

[9] J.A. Tropp, *Topics in Sparse Approximation*, Ph.D. thesis, University of Texas, Austin, 2004.

[10] I. Daubechies, M. Defrise, and C. De Mol, "An iterative thresholding algorithm for linear inverse problems with a sparsity constraint," *Comm. Pure Appl. Math*, vol. 57, pp. 1413–1541, 2004.

[11] D.L. Donoho and I.M. Johnstone, "Ideal spatial adaptation via wavelet shrinkage," *Biometrika*, vol. 81, no. 3, pp. 425–455, 1994.

[12] L. Landweber, "An iterative formula for Fredholm integral equations of the first kind," *Americam Journal of Mathematics*, vol. 73, pp. 615–624, 1951.

[13] M. Elad, "Why simple shrinkage is still relevant for redundant representations?," to appear in IEEE Trans. on Information Theory.