RC24814 (W0906-104) June 23, 2009
Computer Science

IBM Research Report

Lossy Speech Compression Via Compressed Sensing-Based Kalman Filtering

Avishy Carmi
Signal Processing and Communications Laboratory
University of Cambridge
UK

Dimitri Kanevsky, Bhuvana Ramabhadran
IBM Research Division
Thomas J. Watson Research Center
P.O. Box 218
Yorktown Heights, NY 10598
USA

Research Division
Almaden - Austin - Beijing - Cambridge - Haifa - India - T. J. Watson - Tokyo - Zurich

Lossy Speech Compression Via Compressed Sensing-Based Kalman Filtering

Avishy Carmi[1], Dimitri Kanevsky[2] and Bhuvana Ramabhadran[2]

[1]Signal Processing and Communications Laboratory, University Of Cambridge, UK.
[2]IBM T. J. Watson Research Center, Yorktown, NY 10598, USA

Abstract

We present a new algorithm for lossy speech compression. The new algorithm is based on a simple technique for embedding a compressed sensing mechanism within a conventional Kalman filter. As such, it is capable of constructing compressed representations using significantly less samples than what is usually considered necessary.

Index Terms: Lossy Compression, Compressed Sensing, Speech coding

1. Introduction

Recent studies have shown that sparse signals can be recovered accurately using less observations than what is considered necessary by the Nyquist/Shannon sampling principle; the emergent theory that brought this insight into being is known as compressed sensing (CS) [1, 2]. The essence of the new theory builds upon a new data acquisition formalism, in which compression plays a fundamental role. From a filtering standpoint, one can think about a procedure in which signal recovery and compression are carried out simultaneously, thereby reducing the amount of required observations. Sparse, and more generally, compressible signals arise naturally in many fields of science and engineering. A typical example is the reconstruction of images from under-sampled Fourier data as encountered in radiology, biomedical imaging and astronomy. Other applications consider model-reduction methods to enforce sparseness for preventing over-fitting and for reducing computational complexity and storage capacities. The reader is referred to the seminal work reported in [2] and [1] for an extensive overview of the CS theory.

The CS paradigm was used for audio compression in [3]. Relying on some classical techniques for solving the CS problem, such as basis pursuit and orthogonal matching pursuit, the method derived in [3] constructs a sparse discrete cosine transform representation of the underlying audio signal.

In this work we utilize a newly developed CS-based Kalman-filtering algorithm for obtaining a compressed frequency-domain representation of a speech signal from under-sampled time sequence. **In virtue of its CS mechanism, the new algorithm, which is termed CSKF in [4], reconstructs an approximate short-time discrete Fourier transform (DFT) of the signal using significantly less samples than conventional methods. The method derived here shows the potential of the CS approach in reducing the amount of acquired samples in parametric audio compression methods such as harmonic and individual lines and noise (HILN) algorithm [5].**

This paper is organized as follows. The next section briefly reviews the CS approach in the framework of linear estimation. Section 3 presents the CSKF algorithm. The application of the CSKF to lossy speech compression is discussed in Section 4. Finally, a numerical example and conclusions follow thereafter.

2. Sparse Linear Estimation

Consider an $\mathbb{R}^n$-valued random discrete-time process $\{x_k\}_{k=1}^{\infty}$ that is sparse in some known orthonormal sparsity basis $\psi \in \mathbb{R}^{n\times n}$, that is

$$z_k = \psi^T x_k, \quad \#\{\mathrm{supp}(z_k)\} < n \tag{1}$$

where $\mathrm{supp}(z_k)$ and $\#$ denote the support of z_k and the cardinality of a set, respectively. Assume that z_k evolves according to

$$z_{k+1} = Az_k + w_k, \quad z_0 \sim \mathcal{N}(\mu_0, P_0) \tag{2}$$

where $A \in \mathbb{R}^{n\times n}$ and $\{w_k\}_{k=1}^{\infty}$ is a zero-mean white Gaussian sequence with covariance $Q_k \geq 0$. Note that (2) does not necessarily imply a change in the support of the signal. For example, A can be a block-diagonal matrix decomposed of A^d and A^n corresponding to the statistically independent elements $z^d \notin \mathrm{supp}(z_k)$ and $z^n \in \mathrm{supp}(z_k)$ where the respective noise covariance sub-matrices satisfy $Q^d = 0$ and $Q^n \geq 0$. The process x_k is measured by the $\mathbb{R}^m$-valued random process

$$y_k = Hx_k + \zeta_k = H'z_k + \zeta_k \tag{3}$$

where $\{\zeta_k\}_{k=1}^{\infty}$ is a zero-mean white Gaussian sequence with covariance $R_k > 0$, and $H := H'\psi^T \in \mathbb{R}^{m\times n}$ is a sensing matrix.

Letting $y^k := [y_1, \ldots, y_k]$, our problem is defined as follows. We are interested in finding a y^k-measurable estimator, $\hat{x}_k$, that is optimal in some sense. Often, the sought after estimator is the one that minimizes the mean square error (MSE) $E\left[\| x_k - \hat{x}_k \|_2^2\right]$. It is well-known that if the linear system (2), (3) is observable then the solution to this problem can be obtained using the Kalman filter (KF). On the other hand, if the system is unobservable, then the regular KF algorithm is useless; if, for instance, $A = I_{n\times n}$, then it may seem hopeless to reconstruct x_k from an under-determined system in which $m < n$ and $\mathrm{rank}(H) < n$. Surprisingly, this problem may be circumvented by taking into account the fact that z_k is sparse.

2.1. Compressed Sensing

Refs. [1, 2] have shown that in the deterministic case (i. e., when z is a parameter vector), one can accurately recover z (and therefore also x, i.e., $x = \psi z$) by solving the optimization problem

$$\min \| \hat{z} \|_0 \text{ s.t. } \sum_{i=1}^{k} \| y_i - H'\hat{z} \|_2^2 \leq \epsilon \qquad (4)$$

for a sufficiently small ϵ, where

$$\| v \|_p = \left(\sum_{j=1}^{n} v_j^p \right)^{1/p} \qquad (5)$$

is the l_p-norm of v, and the zero-norm, $\| v \|_0$, is defined as[1] $\| v \|_0 := \#\{\mathrm{supp}(v)\}$.

Following a similar rationale, in the stochastic case the sought-after optimal estimator satisfies [6]

$$\min \| \hat{z}_k \|_0 \text{ s.t. } E_{z_k|y^k}\left[\| z_k - \hat{z}_k \|_2^2\right] \leq \epsilon \qquad (6)$$

Unfortunately, the above optimization problems are NP-hard and cannot be solved efficiently. Recently, it has been shown that if the sensing matrix H' obeys a so-called *restricted isometry property* (RIP) while z is sparse enough possibly with [6]

$$s = \mathcal{O}(m/\log(n/m)) \qquad (7)$$

where $s = \#\{\mathrm{supp}(z)\}$, then the solution of the combinatorial problem (4) can almost always be obtained by solving the constrained convex optimization [1, 6]

$$\min \| \hat{z} \|_1 \text{ s.t. } \sum_{i=1}^{k} \| y_i - H'\hat{z} \|_2^2 \leq \epsilon \qquad (8)$$

[1] For $0 \leq p < 1$, $\| v \|_p$ is not a norm; the common terminology is *zero norm* for $p = 0$ and *quasi-norm* for $0 < p < 1$.

This is a fundamental result in the new emerging theory of compressed sensing (CS) [1, 6]. The main idea is that the convex l_1 minimization problem can be efficiently solved using a myriad of existing methods, such as LASSO [7], the Dantzig selector [8], Basis pursuit and Basis pursuit de-noising [9], to mention only a few.

3. The CSKF

For the system described by (2) and (3) the classical KF provides an estimate $\hat{z}_k$ that is a solution to the unconstrained l_2 minimization problem

$$\min_{\hat{z}_k} E_{z_k|y^k}\left[\| z_k - \hat{z}_k \|_2^2\right]$$

Inspired by the CS approach while retaining the KF objective function, we replace (6) by the dual constrained optimization [4]

$$\min_{\hat{z}_k} E_{z_k|y^k}\left[\| z_k - \hat{z}_k \|_2^2\right] \text{ s.t. } \| \hat{z}_k \|_1 \leq \epsilon' \qquad (9)$$

The constrained optimization problem (9) can be solved in the framework of Kalman filtering using the pseudo-measurement (PM) technique [4]. The idea is fairly simple: the inequality constraint $\| z_k \|_1 \leq \epsilon'$ is incorporated into the filtering process using a fictitious measurement $0 = \| z_k \|_1 - \epsilon'$, where ϵ' serves as a measurement noise. This PM can be rewritten as

$$0 = \bar{H} z_k - \epsilon', \quad \bar{H} := [\mathrm{sign}(z_k(1)), \ldots, \mathrm{sign}(z_k(n))] \qquad (10)$$

where $\mathrm{sign}(z_k(i))$ denotes the sign function of the ith element of z_k (i.e., $\mathrm{sign}(z_k(i)) = 1$ if $z_k(i) > 0$ and equals 0 otherwise) . In this setting, the covariance R_ϵ of ϵ' is regarded as a tuning parameter, which can be determined based on simulation runs. A single iteration of the CS-embedded KF is detailed in Algorithm 1 [2].

4. Lossy Speech Compression

Speech is a compressible signal. Usually, vowels can be represented using a limited number of frequencies for which the human hear is most sensitive. The cardinality of this set of significant frequencies may serve as an analog measure to sparseness degree $\#\{\mathrm{supp}(z)\}$. A more formal argument proceeds as follows.

Let $z \in \mathbb{R}^n$ be the DFT of y_k over the discrete times $k = 1, \ldots, n$, that is

$$z(j) = \sqrt{n}^{-1} \sum_{k=1}^{n} y_k \exp\left(-\frac{2\pi i}{n}(j-1)(k-1)\right) \qquad (14)$$

[2] Notice that this is an unusual implementation of the KF as the matrix $\bar{H}_\tau$ is state dependent.

Algorithm 1 CSKF-1 [4]

1: *Prediction*

$$\hat{z}_{k+1|k} = A\hat{z}_{k|k} \tag{11a}$$

$$P_{k+1|k} = AP_{k|k}A^T + Q_k \tag{11b}$$

2: *Measurement Update*

$$K_k = P_{k+1|k}H'^T\left(H'P_{k+1|k}H'^T + R_k\right)^{-1} \tag{12a}$$

$$\hat{z}_{k+1|k+1} = \hat{z}_{k+1|k} + K_k\left(y_k - H'\hat{z}_{k+1|k}\right) \tag{12b}$$

$$P_{k+1|k+1} = (I - K_kH')P_{k+1|k} \tag{12c}$$

3: *CS Pseudo Measurement:* Let $P^1 = P_{k+1|k+1}$ and $\hat{z}^1 = \hat{z}_{k+1|k+1}$.

4: **for** $\tau = 1, 2, \ldots, N_\tau - 1$ iterations **do**

5:

$$\bar{H}_\tau = [\text{sign}(\hat{z}^\tau(1)), \ldots, \text{sign}(\hat{z}^\tau(n))] \tag{13a}$$

$$K^\tau = P^\tau\bar{H}_\tau^T\left(\bar{H}_\tau P^\tau\bar{H}_\tau^T + R_\epsilon\right)^{-1} \tag{13b}$$

$$\hat{z}^{\tau+1} = (I - K^\tau\bar{H}_\tau)\hat{z}^\tau \tag{13c}$$

$$P^{\tau+1} = (I - K^\tau\bar{H}_\tau)P^\tau \tag{13d}$$

6: **end for**

7: Set $P_{k+1|k+1} = P^{N_\tau}$ and $\hat{z}_{k+1|k+1} = \hat{z}^{N_\tau}$.

for $j = 1, \ldots, n$, which can be compactly written as

$$z = \mathcal{F}y \tag{15}$$

where $\mathcal{F}$ and $y \in \mathbb{R}^n$ denote the DFT matrix and a vector whose components are the time points y_j, respectively. Denote F_ϵ the set of ϵ-significant frequencies, and let

$$F_\epsilon = \{z(j) \mid \quad 10\log|z(j)| > \epsilon\} \tag{16}$$

that is, all frequencies for which the amplitude is greater than ϵ dB. Following this definition, $\#F_\epsilon$ is an analog measure to sparseness degree where $n/\#F_\epsilon$ is the compression ratio.

In this work we use the CSKF for reconstructing a frequency representation z of a speech signal from under-sampled time series. In other words, our reconstruction algorithm solves the following problem

$$y = \mathcal{F}_m^* z + \zeta, \qquad y \in \mathbb{R}^m, \quad m < n \tag{17}$$

where $\mathcal{F}_m^* \in \mathbb{R}^{m\times n}$ denotes a sub-matrix obtained by sampling m rows from the inverse DFT matrix (which, in this case, is the conjugate transpose of $\mathcal{F}$). If we follow the arguments presented in [6] (Theorem 2.1) for sparse signals, we may say that in this case an adequate frequency representation is highly probable provided that

$$m \geq c \cdot \#F_\epsilon \log n \tag{18}$$

5. Numerical Study

The CSKF was for reconstructing the short time DFT of a speech recording from a series of overlapping Hamming windows. The algorithm utilized $N_\tau = 200$ PM iterations with $R_\epsilon = 100^2$ and $\alpha = 1$. The window size was set to 256 with only 6 non-overlapping elements. In this example, our DFT vector z is composed out of $n = 256$ elements corresponding to the amplitude and phase of 128 frequencies. Taking the frequency threshold parameter $\epsilon = 0$ in (16) yields $\#F_\epsilon$ between 10 to 20 for the specific signal considered. A rough estimate based on (18) suggests that we need around $m = 110c$ samples picked at each time window for a 'good' frequency representation. We have tested the algorithm with $m = 165$ samples, i.e., the algorithm uses 65% of the available data. The results of this experiment are summarized in Figs. 1 and 3.

5.0.1. Results

The entire time series is shown in Fig. 2a. A typical random sampling pattern when using 65% of the samples in a single time window is shown in Fig. 2b. The original short time DFT of the signal (i.e., when using all available data) is depicted via a spectrogram in Fig. 1a. The reconstructed short time DFT based on the under-sampled data is shown in Fig. 1b. The 128 original (dotted line) and reconstructed (solid line) amplitudes of Fourier coefficients at a single time point are shown in Fig. 3.

6. Conclusions

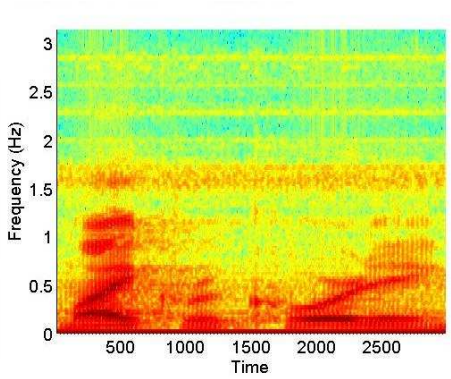
A new lossy compression algorithm is presented. By embedding a compressed sensing mechanism within a conventional Kalman filter, the new algorithm is capable of obtaining an approximate frequency-domain representation of the speech signal using under-sampled time sequence. The method derived here shows the potential of the compressed sensing approach in reducing the amount of acquired samples in parametric audio compression methods.

An extended version of this algorithm is currently being developed for introducing smoother transitions between windows using autoregressive models.

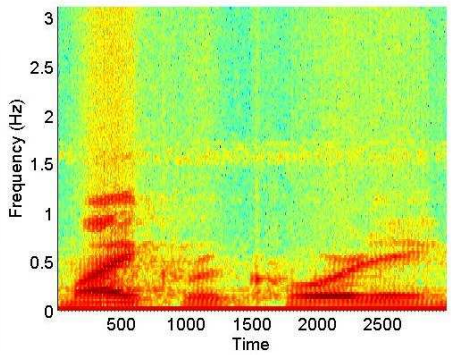
7. References

[1] E. J. Candes, J. Romberg, and T. Tao, "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information," *IEEE Transactions on Information Theory*, vol. 52, pp. 489–509, 2006.

[2] D. Donoho, "Compressed sensing," *IEEE Trans-*

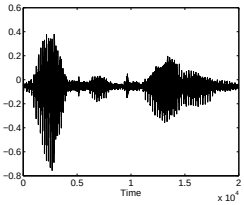


(a) Original 100%



(b) Reconstructed 65%

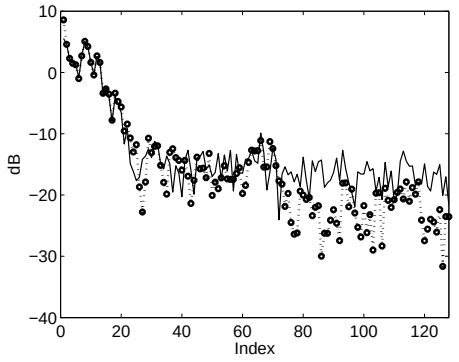
Figure 1: Original and reconstructed short time DFT of a speech signal.



(a) Time domain

(b) Sampling window

Figure 2: The time domain signal and a typical sampling window.



(a) 65%

Figure 3: Actual (dotted line) and recovered (solid line) amplitudes of Fourier coefficients.

actions on Information Theory, vol. 52, pp. 1289–1306, 2006.

[3] A. Griffin and P. Tsakalides, "Compressed sensing of audio signals using multiple sensors." Lausanne, Switzerland: Proceedings of the 16th European Signal Processing Conference (EUSIPCO '08), August 2008.

[4] A. Carmi, P. Gurfil, and D. Kanevsky, "A simple method for sparse signal recovery from noisy observations using kalman filtering," Human Language Technologies, IBM, Tech. Rep. RC24709, 2008.

[5] H. Purnhagen, "Parameter estimation and tracking for time-varying sinusoids." Leuven, Belgium: Proceeding of the 1st IEEE Benelux Workshop on Model based Processing and Coding of Audio (MPCA-2002), 2002.

[6] E. J. Candes, "Compressive sampling," European Mathematical Society. Madrid, Spain: Proceedings of the International Congress of Mathematicians, 2006.

[7] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 58, no. 1, pp. 267–288, 1996.

[8] E. Candes and T. Tao, "The dantzig selector: statistical estimation when p is much larger than n," *Annals of Statistics*, vol. 35, pp. 2313–2351, 2007.

[9] S. S. Chen, D. L. Donoho, and M. A. Saunders, "Atomic decomposition by basis pursuit," *SIAM Journal of Scientific Computing*, vol. 20, no. 1, pp. 33 – 61, 1998.