An Incremental Algorithm for Signal Reconstruction from Short-Time Fourier Transform Magnitude

Jake Bouvrie and Tony Ezzat

Center for Biological and Computational Learning & Department of Brain and Cognitive Sciences
Massachusetts Institute of Technology, Cambridge, MA USA

jvb@mit.edu, tonebone@mit.edu

Abstract

We present an algorithm for reconstructing a time-domain signal from the magnitude of a short-time Fourier transform (STFT). In contrast to existing algorithms based on alternating projections, we offer a novel approach involving numerical root-finding combined with explicit smoothness assumptions. Our technique produces high-quality reconstructions that have lower signal-to-noise ratios when compared to other existing algorithms. If there is little redundancy in the given STFT, in particular, the algorithm can produce signals which also sound significantly better perceptually, as compared to existing work.

Index Terms: signal reconstruction, phase retrieval.

1. Introduction

Reconstruction of a time-domain signal from only the magnitude of the short-time Fourier transform (STFT) is a common problem in speech and signal processing. Many applications, including time-scale modification, speech morphing, and spectral signal enhancement involve manipulating the STFT magnitude, but do not clearly specify how to adjust the phase component of the STFT in order to invert back into the time domain. Indeed, for many STFT magnitude modifications, a valid inverse of the STFT does not exist and a reasonable guess must be made instead.

In this paper, we present an algorithm for reconstruction of a time-domain signal from the STFT magnitude, modified or otherwise. In contrast to existing algorithms based on alternating projections, our technique applies numerical root-finding combined with explicit smoothness assumptions to give high-quality reconstructions. We have found that imposing smoothness at several stages of the algorithm is the critical component responsible for estimating good signals. Formulating the reconstruction problem in terms of non-linear systems of equations serves as a convenient vehicle for the inclusion of smoothness constraints in a straightforward manner. Our method produces results that appear to be perceptually superior to the algorithms due to Griffin and Lim [6] and Achan et al. [1], particularly when there is little overlap between STFT analysis windows.

In section 2 we give an overview of the signal reconstruction problem, and in section 3 we introduce the root-finding framework we have used to find solutions to this problem. Section 4 presents the smoothness constraints we have chosen to impose, followed by a description of the algorithm itself. In section 5 we compare the performance of our technique to Griffin and Lim's method over a range of STFT window overlaps. Finally, in section 6 we offer concluding remarks.

2. Overview of the Phaseless Signal Reconstruction Problem

If the zeros of the Z-transform of a signal lie either entirely inside or outside the unit circle, then the signal's phase may be uniquely related to its magnitude via the Hilbert transform [9]. In the case of finite speech or music signals, however, such a condition on the zeros does not ordinarily hold. Under some conditions, mixed-phase signals can be recovered to within a scale factor from only magnitude or phase [7], and can be uniquely specified from the signed-magnitude [10]. But the conditions required in [7] are restrictive, while retaining any phase-information, albeit even a single bit, is not possible for many common spectrogram modifications.

In this paper, we will focus on reconstruction from magnitude spectra only. Generally, we would like to take a signal, manipulate its magnitude, and from the modified spectra be able to estimate the best possible corresponding time-domain sequence. In the absence of any modifications, we would hope to retrieve the original time-domain signal from the magnitude. If only the Discrete Fourier Transform (DFT) magnitude of a signal is provided, then we must make additional a priori assumptions in order to guess the corresponding signal. This is a common problem in several fields, such as x-ray crystallography, electron diffraction, and remote sensing [5]. If, however, we work with the short-time Fourier transform (STFT), accurate reconstruction is often possible without *a priori* assumptions or constraints. Given a suitable length N windowing function $w(n)$, we can define the STFT by sliding the signal $x(n)$ through the window and taking the K-point DFT:

$$S(\omega_k, \ell) = \sum_{n=0}^{N-1} x(n+\ell)w(n)e^{-j\omega_k n} \tag{1}$$

where the DFT frequency bins are $\omega_k = \frac{2\pi k}{NT}, k = 0, \ldots, K-1$ given sampling rate $f_s = 1/T$. Because both the magnitude $|S(\omega_k, \ell)|$ and phase $e^{j\phi(\omega_k,\ell)}$ of the STFT contain information about the amplitude and phase of the original signal, throwing away the STFT phase does not mean that we have entirely eliminated the original phase of $x(n)$ [4].

Several algorithms have been proposed to estimate a signal from the STFT magnitude. Achan et al. [1] introduced a generative approach for speech signals that infers a time-domain signal from a model trained on a specific speaker or class of speakers. Griffin and Lim [6] apply an iterative technique similar in spirit to an earlier algorithm advanced by Fienup [5]. While it is difficult to analyze the convergence and uniqueness properties of Fienup's algorithm, Griffin and Lim's approach employs alternating convex projections between the time-domain and the STFT domain

that have been shown to monotonically decrease the squared error between the given STFT magnitude and the magnitude of an estimated time-domain signal. In the process, the algorithm thus produces an estimate of the STFT phase. Nawab et al. [8] proposed a sequential algorithm which reconstructs a signal from its spectral magnitude by extrapolating from the autocorrelation functions of each short-time segment, however the approach places sparseness restrictions on the signal and requires that the first h samples of the signal be known, where h is the STFT window hop size. The algorithm presented herein requires neither samples of the signal to be reconstructed, nor does it place constraints on the number of consecutive zeros that can appear in the reconstruction.

3. Signal Reconstruction as a Root-Finding Problem

Griffin and Lim's algorithm attempts to estimate a signal that is consistent with a given spectrogram by inverting the full STFT at each iteration. Alternatively, we can analyze consistency on a column-wise basis, where the spectrogram $|S(\omega_k, \ell)|$ is viewed as a matrix with frequency spanning the rows, and time the columns. Given a single column ℓ_0 from the magnitude of the STFT, we wish to determine the segment of signal $\tilde{x}(n) = x(n + \ell_0)w(n)$ that satisfies the system of equations given by (1):

$$|S(\omega_k, \ell_0)| = \left| \sum_{n=0}^{N-1} \tilde{x}(n) e^{-j\omega_k n} \right|, \quad k = 0, \ldots, K-1. \quad (2)$$

In the discussion that follows, we will abbreviate the above system with the notation $|F\tilde{x}| = m$, where F is the $K \times N$ Fourier matrix, and $m \geq 0$ is the given spectrogram column. Note that although (2) appears to be a system of K equations in N unknowns as it is written, the Fourier magnitude is an even-symmetric function because we allow only real-valued time-domain signals. Thus we really only have $K/2$ linearly independent equations, and 2x oversampling in the DFT is needed to make the system square. In practice we set $K = 2N$ when computing the original STFT, and solve for $\tilde{x}$ using only half of the desired magnitude vector m and a truncated Fourier matrix F. Finally, if we rearrange (2) to get $G(\tilde{x}) \equiv |F\tilde{x}| - m = 0$, $\tilde{x}$ is seen as a root of the function $G : \mathbb{R}^N \to \mathbb{R}^N$ so that estimating the signal is equivalent to solving a numerical root-finding problem.

It should be noted, however, that there are almost always an infinite number of possible roots $\tilde{x}$ satisfying $|F\tilde{x}| - m = 0$, since we can at best match just the magnitude spectra m. Writing $F\tilde{x} = Dm$ in terms of the phasor matrix $D = \text{diag}(e^{j\phi(\omega_k)})$, the phases $\phi(\omega_k)$ need only satisfy the condition $\text{Im}\{F^{-1}Dm\} = 0$. Which root the iteration actually returns will strongly depend on the initial condition $\tilde{x}_0$.

3.1. Solution of non-square systems of nonlinear equations

As we will discuss below, our algorithm involves solving for only a *subset* of the samples in a segment of the signal, while holding the remaining points fixed. One way to solve a system of p nonlinear equations in q unknowns when $p > q$ is to formulate the task as a locally linear least-squares problem. In particular, given a system of equations $f(x) = 0$, suppose that we choose the objective function $\frac{1}{2}\|f(x)\|_2^2$, and linearize $f(x)$ via a Taylor expansion about the point x_k. Defining the Jacobian $J_{ij}(x) = \frac{\partial f_i}{\partial x_j}$, we have

$$\tilde{f}(x) = f(x_k) + J(x_k)(x - x_k). \quad (3)$$

After substituting $\tilde{f}(x)$ into our original objective we arrive at the linear minimization problem

$$x_{k+1} = \underset{x \in \mathbb{R}^q}{\text{argmin}} \{f(x_k)^T f(x_k) + 2(x - x_k)^T J(x_k)^T f(x_k) + (x - x_k)^T J(x_k)^T J(x_k)(x - x_k)\}. \quad (4)$$

Taking the derivative and setting it equal to zero gives a recursive definition for the next point in terms of the current one:

$$x_{k+1} = x_k - \left(J(x_k)^T J(x_k)\right)^{-1} J(x_k)^T f(x_k). \quad (5)$$

Given an initial guess x_0, equation (5) is seen as the classic Gauss-Newton method [2] for the solution of nonlinear least-squares problems.

In practice, one rarely provides closed form expressions for the Jacobian, nor do we want to directly evaluate all p^2 partial derivatives. In fact, for the system $|F\tilde{x}| - m = 0$, the derivative $\frac{d|z|}{dz}$, which shows up in the chain of derivatives needed to compute the Jacobian, does not exist in the sense that the Cauchy-Riemann equations cannot be satisfied; the function $f(z) = |z|$ is not analytic in the complex plane $\mathbb{C}$. We therefore use a variant of Broyden's method [3] in order efficiently compute a numerical approximation to the Jacobian during the course of the iteration (5).

4. Incremental Reconstruction with Regularization

If the STFT (1) is computed with overlapping windows, as is often the case, we can exploit this redundancy in order to estimate a signal from the spectrogram. Both Griffin and Lim's algorithm and the algorithm presented here utilize the constraints on the signal imposed by the overlapping regions when estimating a sequence consistent with the given STFT magnitude. While Griffin and Lim encode these constraints in the form of intersecting convex sets, we recast redundancy in the STFT as the first of two *smoothness constraints*. The second constraint imposes smoothness over a single segment only. Combining these constraints, we construct an initial guess $\tilde{x}_0$ for the current signal segment that can be expected to lead to a good final reconstruction via the iteration (5). This process effectively "biases" the root-finding process towards an appropriate solution.

We additionally assume *positivity* in the reconstruction, in order to eliminate phase sign errors. This constraint requires only that we either add a constant factor to the DC elements of the spectrogram before applying the algorithm, or simply work with a non-negative version of the original signal.

4.1. Smoothness Across Segments

By definition, in the region of overlap the window of signal corresponding to the i-th and $(i + 1)$-th columns of the spectrogram must be the same. If we choose to recover only individual windows $\tilde{x}$ of the signal at a time by solving (2), then the above statement implies that the i-th piece of signal ought to factor into the computation of the $(i + 1)$-th window of signal. This feedback process can be thought of as a form of regularization: the current window of signal must look something like the previous one. The structure of the STFT tells us that the segments must not change too much from one time instant to the next. If the amount of overlap between adjacent windows is greater, then there is a better chance that this assumption will hold.

4.2. Smoothness Within Segments

Overlap constraints provide a good deal of information about $x(n)$, however there are still many possible candidate solutions $\hat{x}(n)$ that satisfy the overlap conditions but do not give back anything near the original signal (when it is known). This problem is amplified when the STFT step size h is large. Therefore, in order to further bias the search for a solution towards a good one, we make an additional smoothness assumption in the region of the window where there is no overlap with the previous segment.

In this region, we must form a reasonable guess as to what the signal might look like when constructing an initial condition $\tilde{x}_0$ for the iterative root-finding procedure. We explore two smooth possibilities in the non-overlapping region: linear extrapolation from a leading or trailing subset of the known overlap points, or zero-order hold extrapolation from the last overlap point. Smoothness can be quantified for both of these methods by examining the energy in the first and second derivatives of a signal constructed by concatenating the "fixed" values with the h extrapolated points. If, in the the linear extrapolation case, we find the energy over the entire signal in the first derivative to be E_1, and in the second derivative to be E_2, then it must be true that for zero-order hold with the same fixed portion of the signal, the resulting signal $x_z(n)$ will have energies

$$\|Dx_z(n)\|_2 \leq E_1, \quad \text{and} \quad \|D^2 x_z(n)\|_2 \geq E_2 \tag{6}$$

where D and D^2 denote first and second discrete derivative operators respectively. Linear extrapolation therefore reduces energy in the second derivative, while zero-order hold continuation will give lower energy in the first derivative. Empirically we have found that linear-extrapolation is preferable when the STFT step size h is small compared to the window size (10% of the window width or less), while zero-order hold yields improved results when h is large. Eventually, linear extrapolation may well produce samples far from the known values as we extrapolate away from the known region of the signal. Thus a mixture of the two methods is yet another possibility, where we might extrapolate for a small number of points relative to the window size and sampling rate, and then hold the final value for the remainder of the extrapolation interval.

We impose one final constraint on each segment. After the root-finding iteration has converged to a solution, we set the mean of the result to the value specified by the DC term of the length N segment's Fourier magnitude, $|S(\omega = 0, \ell_0)|/N$.

4.3. Incremental Signal Reconstruction: Forward-Backward Recursions

The algorithm proceeds by stepping through the STFT magnitude, column by column, first in the *forward* direction, and then heading *backwards*. At each segment, a window of signal is estimated and written to a buffer at a position corresponding to that window's position in the original signal. In the forward direction, smoothness across segments is incorporated when computing a recursive solution to (2) for window $(i+1)$, by explicitly fixing points in the region of overlap with window i to the shared values in the solution returned for that segment. Going backwards, we instead fix the overlapping values for segment i to those previously given by segment $(i+1)$. The very first window of signal in the forward pass is computed from an initial guess $\tilde{x}_0$ comprised of random values drawn from the uniform distribution $\mathcal{U}[0, 1]$. The first backwards pass window is computed from the last forward solution. The full reconstruction is then assembled by overlap-adding the individual time-domain segments according to the original STFT hop size.

The forward pass can be thought of as computing an initial estimate of the signal using previously computed segments for guidance. The backward pass then back-propagates information and constraints accumulated in the forward pass, and can be seen to effectively "repair" errors incurred during the forward computations. Empirically, it is often the case that the first few reconstructions in the forward pass tend to be error prone, due to a lingering influence from the random initial starting point used to launch the algorithm. However, the smoothness constraints we have described quickly guide the roots towards the desired signal values.

Although we have thus far discussed only interactions between adjacent segments of the signal, for STFT hop sizes $h < N$ a given segment will both constrain, and be constrained by, many other segments in a columnar region of the spectrogram. Each window of signal can be thought of as a node within a (cyclic, reachable) network, across which constraints may propagate in any direction. In this framework, recovering the full signal $x(n)$ is akin to finding an equilibrium point where all the constraints are satisfied simultaneously. It is possible that a dynamical systems perspective can be used to describe the behavior of this algorithm.

Finally, we have found that repeating the algorithm on a spectrogram derived from a time-reversed version of the original signal can reduce the reconstruction error further. Specifically, averaging the results under the normal and reversed conditions often times will give a SNR lower than either of the individual reconstructions alone.

A concise summary of our algorithm is given in Algorithm 1, where we have assumed that the size $K \times L$ spectrogram has been computed with windows of length N and hop sizes h.

$\tilde{x}^0 = \text{rand}(\mathcal{U}[0, 1])$
for *all spectrogram columns* $m^i,\ i = 1, \ldots, L-1$ **do**
  · Compute the h elements of $\tilde{x}_0$ by extrapolating from the last p overlapping points in $\tilde{x}^i$
  · Let $\tilde{x}_{ol}$ be the $N-h$ points in $\tilde{x}^i$ that will overlap with $\tilde{x}^{i+1}$
  · Compute the solution $\hat{x}$ to $|F\tilde{x}^{i+1}| - m^{i+1} = 0$ using the Gauss-Newton iteration with initial condition $\tilde{x}_0$, while holding the overlap points in $\tilde{x}^{i+1}$ fixed so that $\tilde{x}^{i+1} = [\tilde{x}_{ol}^T\ \hat{x}^T]^T$
  · Set $\tilde{x}^{i+1}$ = $\tilde{x}^{i+1}$ - mean[$\tilde{x}^{i+1}$] + $m^{i+1}(0)/N$
end
· Repeat the previous loop in the reverse direction, over all spectrogram columns $m^i,\ i = L, \ldots, 1$, extrapolating in the opposite direction with $\tilde{x}^{i-1} = [\hat{x}^T\ \tilde{x}_{ol}^T]^T$ where $\tilde{x}_{ol}$ are the points in $\tilde{x}^{i-1}$ that overlap with segment $\tilde{x}^i$.
· Reconstruct $x(n)$ by overlap-adding the segments $\{\tilde{x}^i\}_{i=1}^L$

Algorithm 1: Incremental Signal Reconstruction Algorithm

5. Experiments

We compared the proposed algorithm to Griffin and Lim's technique on both speech and music signals. In order to better balance the computation times of the two methods, our algorithm was applied only once to the signals and we did not include the time-reversed solution. We additionally applied a mixture of extrapolation techniques when forming the initial root-finding guess. A linear model was fit to the leading/trailing $p = \min(20, N-h)$ points, and extrapolated for 5 points. The remaining unknown points in the non-overlapping region were set to the last extrapolated point. The success of our algorithm does not critically depend on these choices. Griffin and Lim's

algorithm was passed uniformly distributed random initial phase guesses $r_i \sim \mathcal{U}(0,1)$, and was run until the relative decrease in ℓ_2 error between the STFT magnitude of the reconstruction and the given magnitude was less than 0.1%. We separately evaluated Griffin and Lim when given both strictly positive signals and zero-mean signals. For both methods, positivity was enforced by working with spectrograms derived from the target signal $x'(n) = x(n) - \min_m[x(m)]$, rather than $x(n)$ itself.

The speech signals we attempted to reconstruct consisted of a male speaker and a female speaker uttering the phrase "Hi Jane", while the music sample consisted of a percussive drum loop with no other instruments or vocals. The latter example is representative of a class of signals that tends to be more difficult to recover due to abrupt, non-smooth transitions and noisy crashes which dominate the structure of the signal. The signals varied in length from 0.75s to 2.2s, were all sampled at 14.7kHz, and were normalized uniformly. In each experiment, we used a 100 sample (6.8ms) square (boxcar) window and 200 FFT bins. The STFT hop size, however, was systematically varied from 10% to 90% of the window width in steps of 10 samples. We then compared the power signal-to-noise ratio (SNR) between the original signal x_o and the reconstruction x_r for each STFT hop size, where

$$\mathcal{SNR} = 20 \log_{10} \left(\frac{\|x_o\|_2}{\|x_o - x_r\|_2} \right) \text{(dB)}. \qquad (7)$$

While we have found that both methods are stable with respect to initial conditions, the experiments were nevertheless repeated several times.

We show the averaged performance, over 200 trials, on the male speech sample for both algorithms as a function of STFT hop size in Figure 1, where the trace denoted "incremental" corresponds to our technique. It can be seen that our algorithm consistently outperforms Griffin and Lim's algorithm as measured by SNR over the full range of hop sizes. At approximately $h = 30$, positivity of the input signal affects Griffin and Lim's performance. Overall, it is evident that our technique degrades more gracefully as redundancy in the STFT is reduced.

While these results are encouraging, Griffin and Lim's algorithm can give perceptually good results even though the SNR is poor. Often times this can be attributed to inaudible sign errors in the reconstruction, particularly for small hop sizes. With larger hop sizes, we have observed that the error is mainly due to poor reconstruction and significant distortion can be heard. For this reason, it is important to compare the perceptual quality of the two algorithms. In most cases our algorithm is perceptually better over the full range of hop sizes, and the distinction is greater as the STFT analysis window size is increased (while maintaining similar hop sizes as a percentage of window width). In an effort to provide a fair comparison, we have made available a web page [11] where the reader can listen to audio comparisons under several conditions, including those described above.

For small STFT hop sizes our algorithm can require more computation time than the Griffin-Lim algorithm, depending on the number of iterations needed to meet the Griffin-Lim convergence criteria. Otherwise, the two algorithms are generally comparable in running-time.

6. Conclusions

The algorithm we have presented typically achieves greater signal-to-noise ratios than existing reconstruction techniques, and the perceptual improvement for speech and music signals is particularly noticeable when there is less redundancy in the STFT.

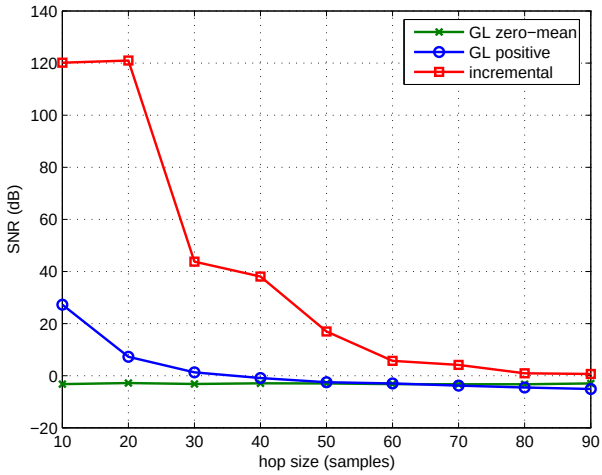

Figure 1: *Algorithm performance (dB) vs. STFT hop size for a 100 sample analysis window*

In designing the algorithm, we imposed several time-domain regularization constraints: (1) We exploited the overlap constraints inherent in the structure of the STFT explicitly and enforced smoothness across windows of the signal. (2) We enforced smoothness within an individual segment by extrapolating in the region where samples were unknown. And, (3) we propagated these constraints throughout the entire signal by applying the smoothness assumptions recursively in both forward and backward directions. We then incorporated these time-domain constraints into a root-finding problem in the frequency domain. Collectively, the constraints can be thought of as biasing the spectral root-finding procedure at each segment towards smooth solutions that are shown to be highly accurate when the true values of the signal are available for comparison.

7. References

[1] K. Achan, S.T. Roweis and B.J. Frey, "Probabilistic Inference of Speech Signals from Phaseless Spectrograms", In S. Thrun et al. (eds.), Advances in Neural Information Processing Systems 16, MIT Press, Cambridge, MA, 2004.

[2] Bertsekas, D., Nonlinear Programming. Athena Scientific, Belmont, MA, 2nd edition, 1999.

[3] Broyden, C. G., "A Class of Methods for Solving Nonlinear Simultaneous Equations", Mathematics of Computation, 19(92):577-593, 1965.

[4] L. Cohen, Time-frequency Analysis, Prentice Hall, Englewood Cliffs, NJ, 1995.

[5] J.R. Fienup, "Phase retrieval algorithms: a comparison", Appl. Opt. 21, pp. 2758-2769, 1982.

[6] D.W. Griffin and J.S. Lim, "Signal reconstruction from short-time Fourier transform magnitude", IEEE Trans. Acoust., Speech, and Signal Proc., 32(2):236-243, 1984.

[7] M.H. Hayes, J.S.Lim, and A.V. Oppenheim, "Signal reconstruction from phase or magnitude", IEEE Trans. Acoust., Speech, and Signal Proc., 28(6):672-680, 1980.

[8] S.H. Nawab, T.F. Quatieri, and J.S. Lim, "Signal reconstruction from short-time Fourier transform magnitude", IEEE Trans. Acoust., Speech, and Signal Proc., 31(4):986-998, 1983.

[9] T.F. Quatieri and A.V. Oppenheim, "Iterative techniques for minimum phase signal reconstruction from phase or magnitude", IEEE Trans. Acoust., Speech, and Signal Proc., 29(6):1187-1193, 1981.

[10] P.L. Van Hove, M.H. Hayes, J.S. Lim, and A.V. Oppenheim, "Signal reconstruction from signed Fourier transform magnitude", IEEE Trans. Acoust., Speech, and Signal Proc., 31(5):1286-1293, 1983.

[11] `http://web.mit.edu/jvb/www/signalrec/`