A novel method for detecting coalescing binaries in near realtime with Advanced LIGO and beyond

Kipp Cannon*
Canadian Institute for Theoretical Astrophysics, 60 St. George Street, University of Toronto, Toronto, ON M5S 3H8, Canada

Adrian Chapman, Antony C. Searle,† and Alan J. Weinstein‡
LIGO Laboratory, California Institute of Technology, Pasadena, CA 91125, USA

Chad Hanna§
Perimeter Institute for Theoretical Physics, Waterloo, Ontario N2L 2Y5, Canada

Drew Keppel¶
Albert-Einstein-Institut, Max-Planck-Institut für Gravitationsphysik, D-30167 Hannover, Germany and Leibniz Universität Hannover, D-30167 Hannover, Germany

(Dated: February 7, 2011)

Conventional matched filter bank methods for the detection of gravitational waves from the inspiral of compact binaries are computationally expensive, have hundreds of seconds of unavoidable intrinsic latency, and require arrays of largely redundant matched filters. Novel detection methods that are more computationally efficient and have lower latency will be required to realize the full potential of advanced of gravitational wave detectors that are currently under construction. In this paper, we describe a new detection method that exploits the properties of inspiral waveforms using multi-rate filtering, principal component analysis, and hierarchical detection. We provide receiver operating characteristics from a prototype search pipeline that is capable of low-latency or near-realtime detection with greatly reduced computational requirements in comparison with previously described methods.

PACS numbers: 95.55.Ym, 84.30.Vn, 95.75.Wx

Submit to PRD or CQG?

* kipp.cannon@ligo.org
† antony.searle@ligo.org
‡ alan.weinstein@ligo.org
§ chad.hanna@ligo.org
¶ drew.keppel@ligo.org

I. INTRODUCTION

In a series of detector upgrades that are now under way, LIGO and Virgo will become the first advanced gravitational wave detectors, gaining a tenfold improvement in amplitude sensitivity [CITATION NEEDED] and a corresponding thousandfold increase in observable volume in the local universe. It is anticipated that these detectors will be able to sense as many as 40 events per year [1].

As the gravitational wave detection horizon pushes outward, the ability to detect signals in near realtime will become valuable. Having an electromagnetic or neutrino counterpart to a gravitational-wave detection would not only increase the confidence in the detection but will also greatly improve the astrophysical information available from the event. Most models that predict simultaneous gravitational-wave and electromagnetic observations also predict that the peak amplitude of electromagetic radiation will occur soon after gravitational-wave emission [2]. Thus in order to maximize the chance of a successful electromagnetic followup the latency of gravitational-wave signal analysis must be made minimal.

Work has already commenced to trigger gravitational-wave searches from electromagnetic observations [3]. LIGO observations ruled out the possibility of GRB070201 originating from the merger of a NSNS or NSBH system in the Andromeda galaxy [4].

This is old news, and doesn't seem relevant to me: EM-directed gravitational wave searches don't benefit from low-latency detection pipelines.

In 2010, LIGO and Virgo completed a period of joint data taking during which several all-sky detection pipelines operated in a low-latency configuration [CITATION NEEDED]. Promising detection candidates were promptly sent for electromagnetic followup [CITATION NEEDED] to several telescopes including Swift, ROTSE, TAROT, and Zadko.

This question has been investigated and it seems at least possible that gravitational-wave detections could provide a region of the sky to prompt electromagnetic observation followups [2]. Much work is currently underway in providing the best possible source localization from gravitational-wave detector networks [5–7]. Collaborations are currently forming to provide infrastructure for the gravitational-wave Astronomers to provide targets of opportunity for electromagnetic astronomers [8]. Most models that predict coincident gravitational-wave and electromagnetic observations also predict that the peak electromagetic fluence will occur soon after gravitational-wave emission [2]. Thus in order to maximize the chance of a successful electromagnetic followup the latency of gravitational-wave signal analysis must be made minimal.

The parameter space of compact binary coalescence signals is large [9, 10]. It is a computationally burdensome task to analyze these signals with even moderate latency [11]. This work will describe how to exploit degeneracy in the signal parameter space to answer more quickly whether or not a gravitational-wave is present. Specifically we will explore using the singular value decomposition (SVD) to reduce the effective number of filters required to search the data. We note that others have applied the use of SVD to gravitational wave data analysis to analyze optimal gravitational-wave burst detection [12, 13] and coherent networks of detectors [14] .

Cite an in-preparation paper of Larry's?

The paper is organized as follows. First we provide an overview of the standard method for detecting compact binary coalescence signals and describe how it can be modified to accomodate low latency analysis. We then describe the pipeline we have constructed to implement these changes. To validate the approach we present results of simulations and finish with some concluding remarks.

II. METHOD

A. The standard approach: matched filtering

Searches for gravitational waves from compact binary coalescences typically employ matched filter banks [15]. Potential inspiral signals are continuously parameterized by time, amplitude, phase, and a set of intrinsic source parameters θ, which in this paper we shall take to consist of the two component masses of a binary, $\theta = (m_1, m_2)$. Let $h_+(\theta)$ and $h_\times(\theta)$ be, respectively, the '+' and '×' polarization gravitational wave signals that would arise from a fiducial face-on binary at some distance. Because, for inspiral signals, h_+ and $h_\times$ are nearly in quadrature, they are generally combined into a single complex-valued template $h = h_+ + ih_\times$.

The detection procedure for just one set of intrinsic source parameters θ starts by *whitening* the measurement data stream x. This involves finding a linear filter that renders the detector's noise IID and Gaussian. This filter is applied to the measurement, yielding the whitened data stream x^{W}. The same linear filter is applied to the template $h(\theta)$, yielding the whitened template $h^{\mathrm{W}}(\theta)$. The matched filter is the normalized cross-correlation of h^{W} and x^{W},

$$\rho(\theta) = \frac{h^{\mathrm{W}}(\theta) \star x^{\mathrm{W}}}{|h^{\mathrm{W}}(\theta)|}.$$

After introducing the ⋆ notation for cross correlation, must we define it for contin-

This is called the signal to noise ratio, or SNR. The detection statistic is the modulus of this, $|\rho(\theta)|$, which has a χ^2 distribution with 2 degrees of freedom in the absence of signal.

To construct a filter bank, matched filters are realized for discrete signal parameters θ_1, θ_2, $\cdots$, θ_N, such that any possible signal will have a maximum cross-correlation of at least 0.97 with at least one template. Such a template bank is said to have a 97% *minimum match*. This technique is designed so that an inspiral signal can be detected without any prior knowledge of its intrinsic parameters: at most 3% of the SNR is lost by a signal's parameters not exactly coinciding with a template's. A trigger is reported for the template parameters θ_i and time t for which $|\rho|$ is a maximum over some moving interval in θ and t.

Citation needed for template placement procedure?

1. Latency and overhead

The matched filter bank can be implemented with finite impulse response (FIR) filters. FIR filters are perfectly suited for realtime detection, because they do not introduce any latency at all. However, FIR filters are very expensive: for N templates of M samples each, a FIR matched filter bank costs $\mathcal{O}(MN)$ operations per sample.

Much more commonly, the matched filter bank is implemented using FFT convolutions, costing only $\mathcal{O}(M \lg N)$ per sample but having a latency of at least M samples. For example, for a bank of 1 ks templates sampled at 4096 Hz, the FFT implementation requires about about 2.2×10^4 times fewer floating point operations per sample than the FIR implementation. However, the FFT implementation has a latency of at least 1 ks.

This presents a dilemma: it seems that low latency detection using the FIR implementation is prohibitively expensive, whereas computational cheap detection with the FFT method comes with minutes to hours of latency.

In the rest of the section, we will describe a detection strategy that makes use of some very general properties of inspiral template banks in order to evaluate a matched filter bank with far lower computational cost than the FIR method and far less latency than the FFT method.

lg is $\log_2$. Explain nomenclature? Switch to $\log_2$?

B. Selectively reducing the sample rate of the data and template waveforms

Our first innovation is based upon the common knowledge that gravitational waveforms from the inspiral stage of compact binary coalescences are chirps: they are quasi-monochromatic and slowly evolving in frequency.

It is possible to strike a balance between the low latency of the FIR filter method and the low overhead of the FFT method by breaking the templates into disjoint time intervals, or *time slices*. The outputs of the filters for all of the time slices are delayed by the appropriate numbers of samples and then summed to produce the SNRs for original templates in the template bank. Each time slice may be implemented using either FIR filters or FFT convolution. If the ith time slice spans the times $[-t^i_{\text{end}}, -t^i_{\text{start}})$ relative to the time of coalescence, then the latency is $\max\left(2(t^i_{\text{end}} - t^i_{\text{start}}) - t^i_{\text{start}}\right)$, where the maximum is taken over all time slices that are implemented using FFT convolution.

We can exploit the time-frequency structure of the templates by processing each time slice at a reduced sample rate. In the quadrupole approximation [16], the frequency of gravitational radiation resulting from the inspiral of two compact objects evolves according to

$$f = \frac{1}{\pi \mathcal{M}} \left[\frac{5}{256} \frac{\mathcal{M}}{-t} \right]^{3/8} \tag{1}$$

where $\mathcal{M}$ is the chirp mass and t is the time relative to the coalescence of the binary [15, 17–19]. Typically the template is truncated at some time before coalescence, corresponding to a finite frequency which is often chosen to be the frequency at the innermost stable circular orbit, $f_{\text{final}} = f_{\text{ISCO}} = 4400\, M_\odot / M$.

Applying the relationship $f/f_{\text{final}} = (t_{\text{final}}/t)^{3/8}$ it is possible to design time slices such that the templates are critically sampled at monotonically decreasing sample rates. The data stream is downsampled to many different sample rates, then filtered in each time slice, then upsampled to a high common sample rate, and summed. In order to work efficiently with radix-2 FFTs, sample rates and sample counts for time slices are both constrained to be powers of 2.

An example time slice design satisfying these constraints for a $1.4 - 1.4\, M_\odot$ is shown in table I below. For this example, the latency for this time slice design is just even if all of the time slices are implemented with the FFT method. This set of time slices will require for pure FFT cross-correlation without time slices, or for the FIR filter method without time slices.

This idea has been demonstrated by the Virgo Collaboration's MBTA pipeline [20, 21], which operated in a low-latency mode during LIGO's sixth science run and Virgo's second science run in 2010.

A similar procedure can be applied for any signal family that is piecewise band-limited, as long as the time-frequency evolution is understood, either symbolically as in equation (1), or numerically.

The time slice boundary notation is a little awkward and different identifiers are used in a few places in the source code. It would be nice to clean that up.

It may be worth pointing out that time-slicing buys you latency even without exploiting time-frequency struc-

TABLE I: Example of critically sampled, power-of-2 time slices for a $1.4 - 1.4\,M_\odot$ template extending from $f_{\rm low} = 10\,{\rm Hz}$ to $f_{\rm ISCO} = 1571\,{\rm Hz}$ with a time frequency structure given by (1).

C. Reducing the number of filters with the singular value decomposition

Our second innovation exploits the fact that the templates in inspiral template banks are, by design, highly correlated. It is possible to greatly reduce the number of matched filters required to achieve a particular minimum match by designing an appropriate set of orthonormal *basis templates*. A purely numerical technique based on the singular value decomposition (SVD) is demonstrated by the authors in [22].

One may regard a bank of N discretely sampled templates of length M samples, $h^{\rm W}(\theta_i; t_j) = [\mathbf{H}]_{ij} = H_{ij}$, as a matrix. The singular value decomposition is an exact factorization that exists for any matrix such that

$$H_{ij} = [\mathbf{V\Sigma U}]_{ij} = \sum_{k=1}^{N} v_{ik}\sigma_k u_{kj}, \tag{2}$$

where $\mathbf{V}$ and $\mathbf{U}$ are both unitary matrices and $\mathbf{\Sigma}$ is a diagonal matrix. For our purposes, we associate the rows of $\mathbf{U}$ with a minimal set of basis templates, which become the kernels of basis filters. These filters give rise to the orthogonal SNRs, $[\rho^\perp]_k = \rho_k^\perp = \mathbf{U}_k \star x^{\rm W}$. The rows of $\mathbf{V\Sigma}$ become reconstruction coefficients that map linear combinations of the orthogonal SNRs from the basis filters back onto SNRs for the original templates of interest: $\rho_i = \sum_k [\mathbf{V\Sigma}]_{ik}\rho_k^\perp$.

In many applications of the SVD, including ours, the matrix can be well approximated by truncating the summation in equation (2) at $L \ll N$:

$$H'_{ij} = \sum_{k=1}^{L} v_{ik}\sigma_k u_{kj} \tag{3}$$

The cumulative sum of squares of the *singular values*, σ_k, measures the Frobenius norm of the approximation, such that

$$\frac{\|\mathbf{H}'\|}{\|\mathbf{H}\|} = \left(\sum_{k=1}^{L} |\sigma_k|^2\right)^{1/2} \left(\sum_{k=1}^{N} |\sigma_k|^2\right)^{-1/2}. \tag{4}$$

This is also called the SVD tolerance. In our application, it relates to how much SNR is lost by discarding $N - L$ of the basis filters with the lowest singular values.

This result differs from other work that models gravitational-wave chirp signals in approximate ways [23–25] by starting with an exact representation of the desired template family and producing a rigorous approximation with a tunable accuracy.

D. Composite detection statistic and hierarchical detection

Although the SVD allows us to reduce the number of matched filters, this comes at the price of having to perform a matrix multiplication by the reconstruction matrix $\mathbf{V\Sigma}$ at every sample. In some circumstances, this matrix multiplication may be more expensive than applying the orthogonal matched filters.

Further speedup may be gained in a hierarchical detection scheme. In general, the orthogonal SNRs alone provide some indication of whether any template in the original template bank is likely to have a large SNR. Consider some composite detection statistic that is a scalar function of the orthogonal SNRs, $\Gamma(\rho_1^\perp, \rho_2^\perp, \cdots, \rho_L^\perp)$. Suppose that we can infer the distribution of Γ for a data stream with the signal present or with the signal absent. Then we may use a threshold crossing of Γ to trigger the conditional application of the expensive reconstruction matrix only during the times when a signal is likely to be present.

The composite detection statistic that we employ is a weighted sum of squares, $\Gamma = \sum_{k=1}^{L} w_k (\rho_k^\perp)^2$, with the particular weights

$$w_k = \frac{\sigma_k^2}{\sigma_k^2 + N/A^2}. \tag{5}$$

Here, A is a desired SNR scale that is set by the analyst. As the authors show in [27], this choice of weights has a better receiver operating characteristic for a signal of SNR $= A$ than some other obvious choices, $w_k = \sigma_k^2$ or $w_k = 1$.

I'm sweeping the complex nature of the templates under the rug here. Is it OK to regard our packing of the real and imaginary parts of the SVD as an implementation detail? It makes the math a little more concise.

This sounds confrontational to me, but these are good citations. Can we tone it down? We should cite [26] here.

We now have a pretty good description of time slices,

III. IMPLEMENTATION AND ANALYSIS

The detection method described above is much more efficient in terms of floating point operations than the traditional matched filter bank method. However, time slices and conditional reconstruction greatly complicate queueing, synchronizing, and bookkeeping of intermediate signals. A low latency implementation capable of recruiting more than one CPU core would be difficult to achieve within the familiar serial programming framework because of the nontrivial time-delay relationships between samples. Due to these complications, we chose to prototype the search using an open source signal processing environment called `GStreamer`. Primarily used for playing, authoring, or streaming media on `Linux` systems, `GStreamer` is an integral component of the popular `Gnome` desktop.

Should I mention `GStreamer` at all?

Citation for `GStreamer`? It would be good to illustrate the layout of this particular template bank: masses spanned, time slice layout . . . Do we need separate "analysis" and "results" sections for this paper?

TABLE II: Operation counts per sample for six different detection methods. The operation counts for LLOID assume a reconstruction duty cycle of 5%. Note that the FIR method with LLOID is almost 10 times faster than the conventional FFT method, despite having substantially lower latency.

IV. RESULTS

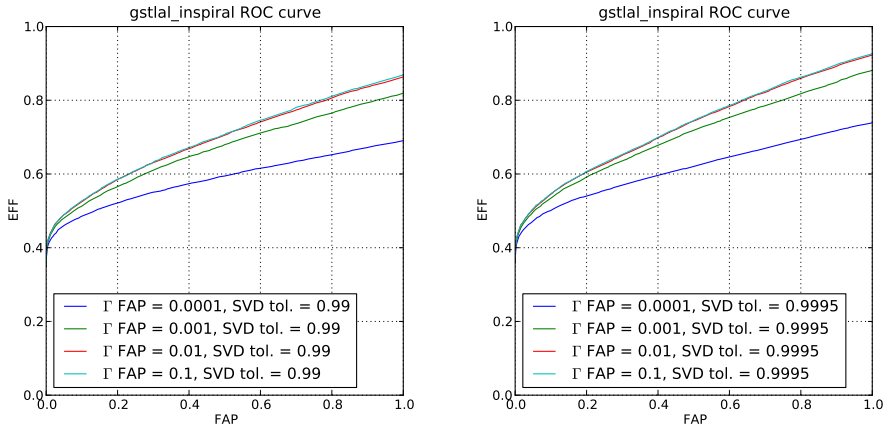

FIG. 1: Receiver operating characteristic (ROC) curve of detection efficiency (EFF) versus false alarm probability (FAP).

V. CONCLUSIONS

ACKNOWLEDGMENTS

LIGO was constructed by the California Institute of Technology and Massachusetts Institute of Technology with funding from the National Science Foundation and operates under cooperative agreement PHY-0107417. This paper has LIGO Document Number LIGO-P0900004-v1.

Appendix A: Floating point operation counts

The filter bank can be implemented using finite impulse response (FIR) filters, which are just sliding window dot

I just yanked this from the methods section.

products. If there are M templates of length n, and the data stream contains N samples, then applying the filter bank requires $2MNn$ operations.

More commonly, the matched filters are implemented using the FFT convolution. This entails applying FFTs to blocks of D samples, with $2n \leq D$, each block overlapping the previous one by n samples. There are $N/(D-n)$ such blocks. Modern implementations of the Cooley-Tukey FFT, such as the ubiquitous `fftw`, require about $4N \lg N$ operations to evaluate a DFT of size N [28]. A D sample cross-correlation consists of a forward FFT, an D sample dot product, and an inverse FFT, totaling $8D \lg N + 2D$ operations per block. Per sample, this is $(8 \lg D + 2)/(1 - n/D)$ operations.

This is more commonly known as "overlap-save". We should find someone else's operation count and cite it.

The FIR filter implementation has the advantage that it has no intrinsic latency, whereas the FFT convolution has at least the latency of the FFT block size $D \geq 2n$. For example, for a $1.4 - 1.4\,M_\odot$ template with duration $\sim 1\,\mathrm{ks}$, the FFT convolution has a latency $\geq 2\,\mathrm{ks}$. However, the FIR filter implementation has the disadvantage of much greater overhead per sample than the FFT convolution. For a 1 ks template sampled at 4096 Hz, the FIR implementation requires about about $n/8 \lg 2n = 2.2 \times 10^4$ times more operations per sample than the FFT implementation.

---

[1] J Abadie, B Abbott, and R Abbott. . . . Predictions for the rates of compact binary coalescences observable by ground-based gravitational-wave detectors. *. . . and Quantum Gravity*, Jan 2010.

[2] Julien Sylvestre. Prospects for the detection of electromagnetic counterparts to gravitational wave events. *Astrophys. J.*, 591(591):1152–1156, 2003.

[3] LIGO Scientific Collaboration and Virgo Collaboration. Astrophysically triggered searches for gravitational waves: Status and prospects. *Class. Quant. Grav.*, 25(25):114051, 2008.

[4] LIGO Scientific Collaboration and K. Hurley. Implications for the origin of GRB 070201 from LIGO observations. *The Astrophysical Journal*, 681(2):1419–1430, 2008.

[5] J. Markowitz, M. Zanolin, L. Cadonati, and E. Katsavounidis. Gravitational wave burst source direction estimation using time and amplitude information. 2008.

[6] V. Raymond, M.V. van der Sluys, I. Mandel, V. Kalogera, C. Roever, and N. Christensen. Degeneracies in sky localisation determination from a spinning coalescing binary through gravitational wave observations: a markov-chain monte-carlo analysis for two detectors. 2008.

[7] F. Cavalier (LAL), M. Barsuglia (LAL), M.-A. Bizouard (LAL), V. Brisson (LAL), A.-C. Clapson (LAL), M. Davier (LAL), P. Hello (LAL), S. Kreckelbergh (LAL), N. Leroy (LAL), and M. Varvella (LAL). Reconstruction of source location in a network of gravitational wave interferometric detectors. *Phys. Rev. D*, 25(25):082004, 2006.

[8] Jonah Kanner, Tracy L. Huard, Szabolcs Marka, David C. Murphy, Jennifer Piscionere, Molly Reed, and Peter Shawhan. LOOC UP: Locating and observing optical counterparts to gravitational wave bursts. 2008.

[9] Benjamin J. Owen. Search templates for gravitational waves from inspiraling binaries: Choice of template spacing. *Phys. Rev. D*, 53:6749–6761, 1996.

[10] Benjamin J. Owen and B. S. Sathyaprakash. Matched filtering of gravitational waves from inspiraling compact binaries: Computational cost and template placement. *Phys. Rev. D*, 60:022002, 1999.

[11] B. Abbott et al. Search for gravitational waves from binary inspirals in S3 and S4 LIGO data. *Phys. Rev.*, D77:062002, 2008.

[12] Patrick R. Brady and Saikat Ray-Majumder. Incorporating information from source simulations into searches for gravitational-wave bursts. *Classical and Quantum Gravity*, 21:S1839–S1848, 2004.

[13] Ik Siong Heng. Supernova waveform catalogue decompositon: a principal component analysis approach. 2008.

[14] Linqing Wen. Data analysis of gravitational waves using a network of detectors. *Int. J. Mod. Phys. D*, 17:1095–1104, 2008.

[15] B. A. Allen, W. G. Anderson, P. R. Brady, D. A. Brown, and J. D. E. Creighton. FINDCHIRP: an algorithm for detection of gravitational waves from inspiraling compact binaries. 2005.

[16] L.S. Finn and D.F. Chernoff. Observing binary inspiral in gravitational radiation: One interferometer. *Phys. Rev. D*, 47:2198–2219, 1993.

[17] Kidder L E, Will C M, and Wiseman A G. Innermost stable orbits for coalescing binary systems of compact objects. *Class. Quantum Grav.*, 9:L125–31, 1992.

[18] Luc Blanchet. Innermost circular orbit of binary black holes at the third post-newtonian approximation. *Phys. Rev. D*, 65, 2002.

[19] Chad Hanna, Miguel Megevand, Evan Ochsner, and Carlos Palenzuela. A method for estimating timefrequency characteristics of compact binary mergers to improve searches for inspiral, merger and ring-down phases separately. *Class. Quantum Grav.*, 26:015009, 2009.

[20] F.Beauville, M.-A.Bizouard, L.Blackburn, L.Bosi, P.Brady, L.Brocco, D.Brown, D.Buskulic, F.Cavalier, S.Chatterji, N.Christensen, A.-C.Clapson, S.Fairhurst, D.Grosjean, G.Guidi, P.Hello, E.Katsavounidis, M.Knight, A.Lazzarini, N.Leroy, F.Marion, B.Mours, F.Ricci, A.Vicere, and M.Zanolin. Benefits of joint LIGO – Virgo coincidence searches for burst and inspiral signals. *J. Phys. Conf. Ser.*, 32(32):212, 2006.

[21] F. Beauville, M.-A. Bizouard, L. Blackburn, L. Bosi, L. Brocco, D. Brown, D. Buskulic, F. Cavalier, S. Chatterji, N. Christensen, A.-C. Clapson, S. Fairhurst, D. Grosjean, G. Guidi, P. Hello, S. Heng, M. Hewitson, E. Katsavounidis, S. Klimenko,

M. Knight, A. Lazzarini, N. Leroy, F. Marion, J. Markowitz, C. Melachrinos, B. Mours, F. Ricci, A. Vicer, I. Yakushin, and M. Zanolin. Detailed comparison of LIGO and Virgo inspiral pipelines in preparation for a joint search. *Class. Quant. Grav.*, 25(25):045001, 2008.

[22] Kipp Cannon, Adrian Chapman, Chad Hanna, Drew Keppel, Antony C Searle, and Alan J Weinstein. Singular value decomposition applied to compact binary coalescence gravitational-wave signals. *Physical Review D*, 82:44025, Aug 2010. (c) :.

[23] Eric Chassande-Mottin and Archana Pai. Best chirplet chain: near-optimal detection of gravitational wave chirps. *Phys. Rev. D*, 73:042003, 2006.

[24] Emmanuel J. Cands, Philip R. Charlton, and Hannes Helgason. Gravitational wave detection using multiscale chirplets. *Class. Quant. Grav*, 25:184020, 2008.

[25] Alessandra Buonanno, Yanbei Chen, and Michele Vallisneri. Detection template families for gravitational waves from the final stages of binary-black-hole inspirals: Nonspinning case. *Phys. Rev. D*, 67:024016, 2003. Erratum-ibid. 74 (2006) 029903(E).

[26] Scharf. Matched subspace detectors. *Signal Processing, IEEE Transactions on*, 42(8):2146 – 2157, 1994.

[27] Kipp Cannon, Chad Hanna, Drew Keppel, and Antony C Searle. Composite gravitational-wave detection of compact binary coalescence. In preparation.

[28] S Johnson and M Frigo. A modified split-radix FFT with fewer arithmetic operations. *Signal Processing, IEEE Transactions on*, 55(1):111 – 119, 2007.