TOWARDS AN AUTOMATED WAVEFORM CORRELATION DETECTOR SYSTEM

Megan E. Slinkard, Dorthe B. Carr, Stephen L. Heck, and Christopher J. Young

Sandia National Laboratories

Sponsored by the National Nuclear Security Administration

Award No. DE-AC04-94AL85000/SL11-WaveCorr-NDD02

ABSTRACT

For nuclear explosion seismic monitoring, major aftershock sequences can be a significant problem because each event must be treated as a possible nuclear test. Fortunately, the high degree of waveform similarity expected within aftershock sequences offers a way to more quickly and robustly process these events than is possible using traditional methods (e.g., short-term averaging/long-term averaging—STA/LTA—detection).

We explore how waveform correlation can be incorporated into an automated event detection system to improve both the timeliness and the quality of the resultant bulletin. With our first iteration of the Waveform Correlation Detector, we processed three aftershock sequences—the 1994 Northridge earthquake, the 2005 Pakistan earthquake, and the 2008 Wenchuan earthquake. Our system compared incoming waveform data with a continuously updating library of known events. Incoming waveform data that correlated above a specified threshold with a library event was marked as a repeating event. We set strict window lengths, filter bands, and correlation thresholds and used them with no regard to the distance between the station and the location of the main shock. Depending upon that distance, between 24% and 92% of the events in a sequence were recognized as repeating events.

We realized early on in the process that the arbitrary window length, filter band, and correlation threshold were not always the best choices for doing waveform correlation. Therefore, we have modified our code to adaptively and in real time determine appropriate parameters for window length, filter bands, and correlation thresholds. We then used this automated waveform correlation detector (WCD) to reprocess some of the datasets from which we had previously obtained poor results; we found the automated WCD improved results considerably, i.e., raising the number of events in a sequence recognized as repeating events from 24% to 47%.

Our system is designed to begin running shortly after a large main shock and adaptively determine parameters appropriate to the swarm as it runs. We use the March 11, 2011, Japan earthquake sequence to demonstrate the automated WCD. We will also present techniques used to improve the processing speed of our algorithm using the Java Parallel Processing Framework, allowing us to perform waveform correlation on larger and larger datasets.

OBJECTIVES

Swarms of earthquakes and/or aftershock sequences can dramatically increase the level of seismicity in a region for a period of time lasting from days to months, depending on the swarm or sequence. For those who monitor seismic events for possible nuclear explosions, these swarms/sequences are a nuisance because each event must be treated as a possible nuclear test until it can be proven, to a high degree of confidence, not to be. Fortunately, swarms typically consist of groups of very similar looking waveforms, suggesting that they can be effectively processed using waveform correlation techniques, which have been shown to have excellent sensitivity and robustness. Our research seeks to use waveform correlation techniques to improve the efficiency of the monitoring system.

We have designed a prototype WCD that is used to simulate applying a waveform correlation-based process to large aftershock sequences. In this manner, we can evaluate the value of these techniques for typical sequences of interest. We used our WCD to simulate using WC techniques on three aftershock sequences of interest: Northridge 1994, Kashmir 2005, and Wenchuan 2008. Our work has demonstrated that a high percentage of events in these sequences could be detected and identified using the WCD. As we move toward an operational system, we must find ways to automate and optimize the selection of parameters and waveforms used by the WCD. In particular, we discuss how to select template waveforms and how to select an appropriate correlation threshold. To avoid confusion, our initial WC Detector used to evaluate the three datasets will henceforth be referred to as our Basic WCD. Our most recent version, which includes our latest automation and optimization techniques, will henceforth be referred to as our Automated WCD. We conclude by showing the improvement resulting from implementing these changes on a typical dataset.

RESEARCH ACCOMPLISHED

Basic WCD

We have designed a prototype WC Detector to simulate using waveform correlation in an operational monitoring system. Waveform correlation techniques rely on the uniqueness of paths through the earth to declare that waveforms with a high degree of correlation are statistically very likely to be from the same source (Thorbjanrardottir et al., 1987; Harris, 1991; Withers et al., 1999). This simple approach is remarkably robust. The high rate of repeated events expected during an aftershock sequence suggests that waveform correlation can identify a large percentage of the aftershocks. In the monitoring mission, where aftershocks are nuisance events, rapid identification of aftershocks improves monitoring efficiency. Although this technique has been known and studied for many years, the computational requirements of calculating correlation have limited its use. Only in the last few years have the computational resources become widely available to quickly process large amounts of data.

Our Basic WCD compares incoming raw data with master waveforms from known events (Figure 1). Our algorithm operates on a single station, during a prescribed time period. The incoming raw data stream is filtered, windowed, and then correlated with each waveform in the Master Waveform Library. If the data stream and a particular library entry have a correlation value above a threshold, we declare a recognized similar event. Detected matches are identified as either a cataloged match (listed in one of the available catalogs) or as a new (uncataloged) event. The incoming data stream is then advanced one sample, and the process repeats.

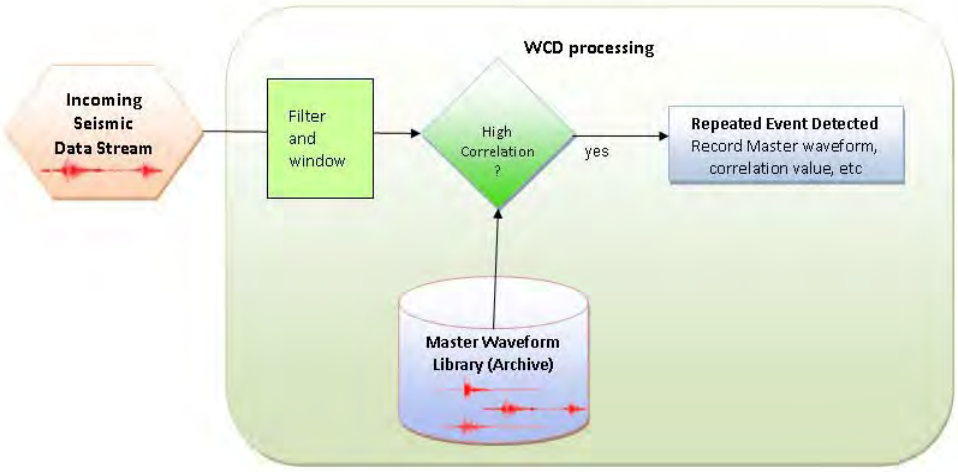


Figure 1. Basic WCD. The incoming raw data stream is filtered, windowed, and then correlated with each waveform in the Master Waveform Library. If a correlation is above the threshold, we say a match is found and record information such as the start time of the data segment, the correlation strength(s), and the master waveform(s) that found the match. The incoming data stream is then advanced one sample, and the process repeats.

Results are written to a database and interpreted as families. A family consists of a master waveform and all of its matches (Figure 2). In this example, the master event found four similar, repeating events in the next two days. All showed similar P and S arrivals. To confirm the correctness of the Family, we can verify that the lat/lon location listings for events in a catalog (EDR in this case) are similar. In an operational setting, we envision an analyst looking at families of similar waveforms, and processing them at the same time.

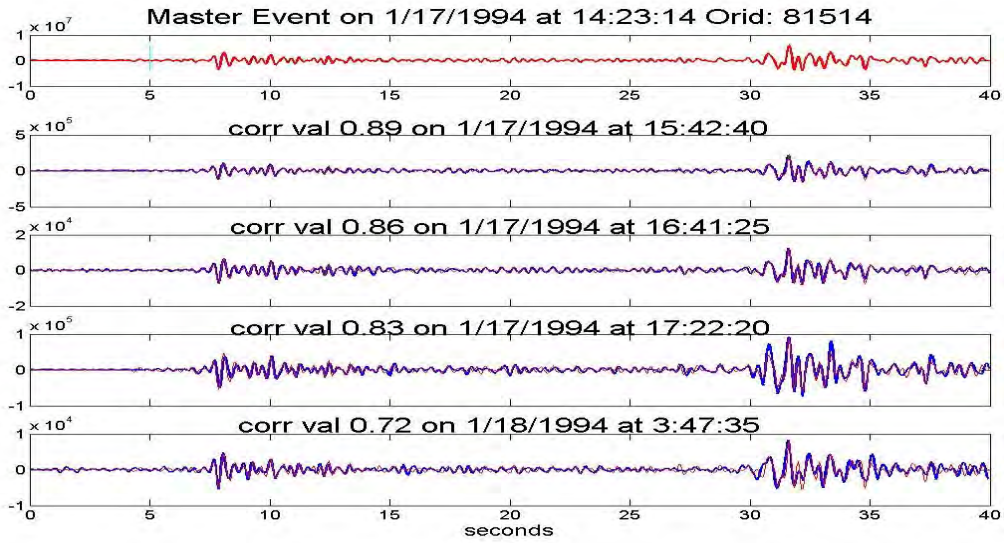


Figure 2. A typical family found by the WCD. The master waveform is shown in red, and the found matches are shown in blue.

Results from Three Datasets

We ran our Basic WCD on three datasets from typical large aftershock sequences: Northridge, Kashmir, and Wenchuan. Each dataset consisted of 5 days of data, starting from just before the mainshock. For each dataset we evaluated the WCD at a near and far station to study the effect of distance on the results. In our Basic WCD runs, the correlation threshold was set to 0.5—this meant events that correlated with a value greater than 0.5 were considered to match. This number was selected to be high enough that we were ensured of matches that an analyst would agree looked similar. Our work turning the threshold selection into an empirical process is described later in this paper. Similarly, the window lengths for the master waveform templates were chosen to be large enough to include both P and S arrivals; the automated version of this process is described later in this paper.

Our results (Table 1) showed that between 17% and 83% of cataloged events belonged to a family of similar waveforms. When we add in families created by matching an uncataloged event to a catalogued event, this percentage range rises to 30%–92%. Moreover, many new, uncataloged signals are identified. Our results suggest that waveform correlation is of value while processing large earthquake aftershock sequences and can dramatically reduce the number of events needing a comprehensive review by an analyst.

Table 1: Results from using the basic WCD on three large aftershock sequences

Event	Northridge	Northridge	Pakistan	Pakistan	Wenchuan	Wenchuan
Station	PAS	MHD	NIL	AAK	CD2	XAN
Station distance (km)	27	348	99	907	39	621
Correlation threshold	0.5	0.5	0.5	0.5	0.5	0.5
Window length (sec)	40	40	40	120	40	90
# catalog events visible at station	412	371	440	360	262	752
% seen catalog events in a family (with other catalog events)	**83%**	**48%**	**51%**	**22%**	**43%**	**17%**
% seen catalog events in a family (with other catalog events OR new signals)	**92%**	**57%**	**78%**	**24%**	**58%**	**30%**
# new signals identified	**942**	**55**	**740**	**10**	**300**	**218**

We then turned our attention to research areas that would need to be explored before an operational system could be implemented. An operational system must do the following:

(1) Recognize a swarm has started
(2) Decide which stations on which to perform WC detection
(3) Determine a library of master waveforms (whether using archived data, incoming data, or both)
(4) Determine a filterband to suitably filter the data
(5) Determine a window length
(6) Determine the correlation threshold(s)
(7) Perform the WC detection
(8) Remove poor-quality matches

Based on our previous work, we determined that window length and correlation threshold had significant influence on the quality of results and needed to be chosen with care. Our goal is twofold: first, to determine optimal parameters, and second, to develop ways of automating the software to determine these optimal parameters.

Window Length

Window length refers to the number of seconds of waveform captured for the master waveforms stored in the Master Waveform Library. It also, therefore, is the number of seconds over which the correlation is performed. We found that this parameter has a significant effect on the number and quality of matches found. Too short a time leads to false matches: S arrivals can correlate with a master waveform's P arrival, short snippets correlate when the overall envelopes don't, etc. A window that is longer than necessary wastes processing time (calculating correlations is computationally expensive) and increases the probability of new arrivals corrupting the signal (in an aftershock sequence, the next event can occur before the shaking from the previous one has subsided). We found that a window length that includes the P arrival and the beginning of the S arrival is the optimal length. This also helps improve accuracy because event to station distance is reflected in the P-S separation.

Given a station and an event region, we determine the difference between theoretical P and S arrivals. We set the window length to 1.2 times the median P to S separation. This allows us to quickly select a constant window length that works well for our master waveforms.

Threshold

Selecting the correlation threshold is one of the most critical factors in the success and accuracy of the WCD. We wanted an objective method for automatically determining a suitable threshold for each master waveform. Calculating the threshold for a given probability of false alarm depends on the time-bandwidth product of the waveform; thus, it depends on the window length and filter band chosen. Using Wychecki-Vergara's technique (Wychecki-Vergara et al., 2001) for determining a threshold, given a suitable probability of error, we originally used station background noise. However, we felt that the threshold returned was too low and that the resultant families did not always look similar to the eye. We decided to instead treat distant events as noise. This method raised the correlation threshold slightly and yielded resultant families that looked similar.

We assume that events more than 50 km distant from each other should not correlate and are effectively noise.

For each master waveform:

- *Compare master to other master events that are >50 km away.*
- *Use Wychecki-Vergara's technique to figure out the threshold for a probability of error that gives 100 years between false matches.*
- *Take the mean of the calculated thresholds to determine the correlation threshold for the master waveform.*

For our Kashmir dataset, this returned correlation thresholds ranging from 0.26 to 0.39, quite a bit lower than our 0.5 starting point, yet higher than using Wycheki-Vergara's method on background noise (0.23 to 0.28).

Data quality

As we experimented with lowering the correlation thresholds, we found that we would occasionally declare a match that an analyst did not agree was a match. This almost always occurred when the incoming data window was very low amplitude except for an arrival at the end. We believe this is an artifact of using the normalized correlation coefficient. To discard these false matches, we check that the energy distribution in the data window is distributed appropriately for a seismic event. If the energy is disproportionately in the last quarter of the window, we discard it (see Figure 3).

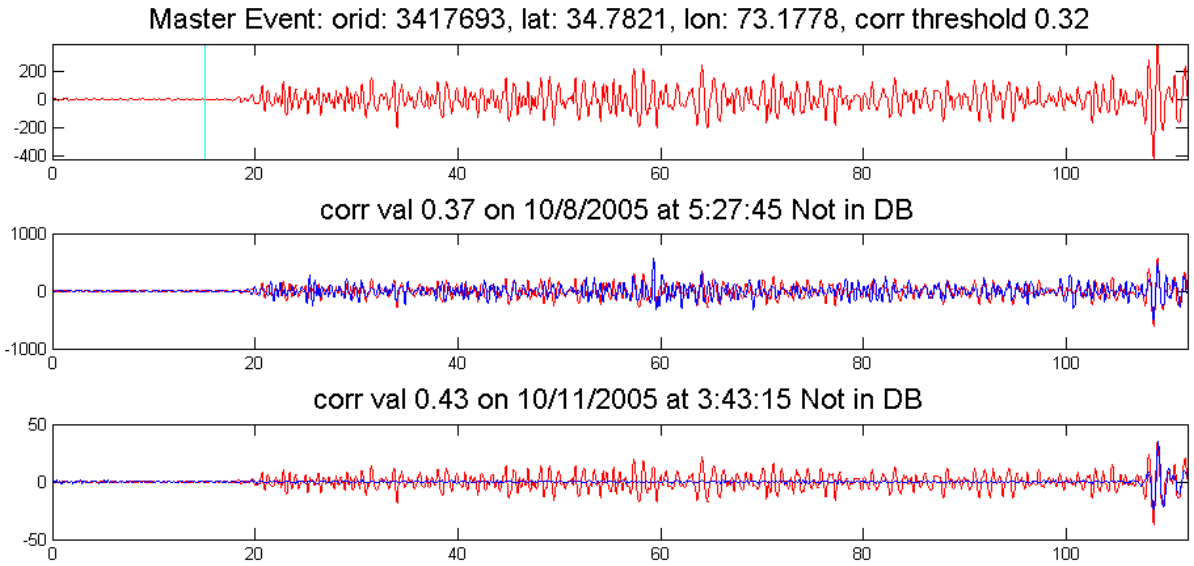


Figure 3. Poor-quality match to be removed. The top plot is the master waveform (red). The middle plot is of a good match, with typical energy distribution. The bottom plot is of a bad match, which will be kicked out by the algorithm.

Automated WCD

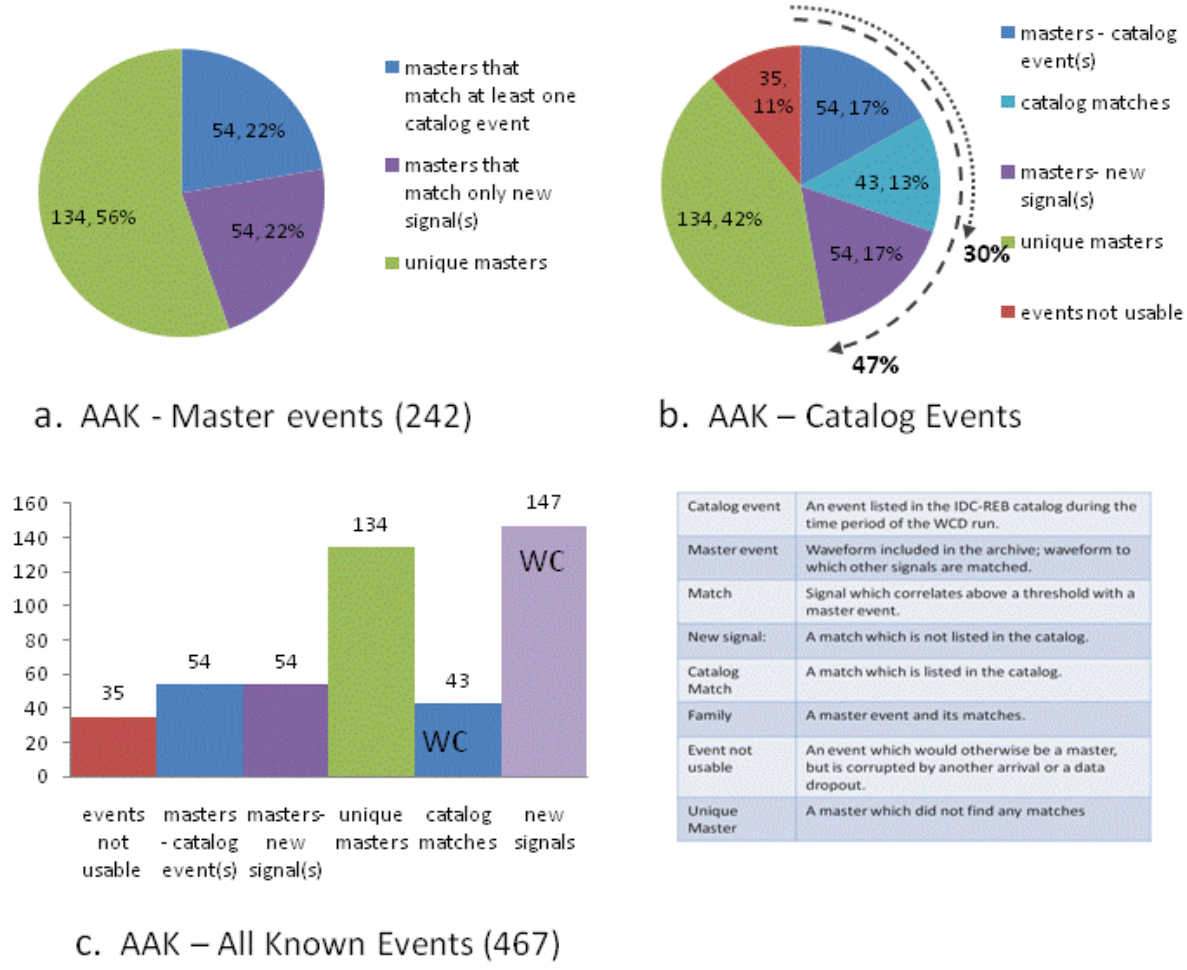
We added the window length, threshold, and data quality code to our WCD to make our Automated WCD. We expect our Automated WCD to continue to evolve as we add and improve methods of automating WC detection. Using the Automated WCD, we reprocessed the worst-performing sequence from our original results: Kashmir at station AAK. Our Automated WCD found many more matches, increasing the percent of catalog events belonging to a family from 24% to 47% and the number of new signals from 10 to 183 (see Table 2).

Table 2. Effectiveness of the WCD on the Kashmir earthquake dataset, with the automated system performing substantially better than the basic system

Run	Station	Filter band	Window length	Corr threshold	# of catalog events seen at station	% catalog events belonging to a family	# of additional events identified
Basic WCD	AAK	0.8–3.5 Hz	120 sec	0.5	360	**24%**	**10**
Automated WCD	**AAK**	**0.8 – 3.5 Hz**	**112 sec (auto)**	**0.26–0.39 (auto)**	**360**	**47%**	**183**

We can look at the results in detail to observe the makeup of the library and families. Figure 4a plots information about the master events used in the master event library: 44% found matches, and 56% did not. The large number of master events that didn't find a match suggests that our master waveform library is unnecessarily large. Future research will explore how to select a smaller, yet still effective library to save analyst and processing time. This is not so important if archival data is used for the library, but it is of high importance if events occurring during the aftershock sequence are to be added to the master event library. Of the master events that did find matches, half found events in the catalog, and half found only new signals (events not in the catalog).

Figure 4b shows the distribution of catalog events. Catalog events are generally either master events or matches. Some events were not usable—this usually refers to events that had multiple arrivals in their window. This chart shows the number of catalog events that belonged to a family of similar events. The dotted line (30%) includes catalog events that matched another catalog event; the dashed line (47%) also includes families that were created using only new signals. Thus, about half of the events in the catalog for AAK belong to a family of similar events.



Catalog event	An event listed in the IDC-REB catalog during the time period of the WCD run.
Master event	Waveform included in the archive; waveform to which other signals are matched.
Match	Signal which correlates above a threshold with a master event.
New signal:	A match which is not listed in the catalog.
Catalog Match	A match which is listed in the catalog.
Family	A master event and its matches.
Event not usable	An event which would otherwise be a master, but is corrupted by another arrival or a data dropout.
Unique Master	A master which did not find any matches

Figure 4. A detailed look at results from the Automated WCD: (a) The breakdown of events considered for the Master Waveform Library. (b) All cataloged events included in the study. (c) The number of new, uncataloged signals found by the detector. Events found by the WCD are marked WC for easy identification.

Figure 4c plots the distribution of all known events, including the 147 new signals. Events found via waveform correlation are marked WC for easy identification. In this scenario 242 library events yielded another 190 events found via waveform correlation.

These plots convey both the value (the percentage of catalog events that belong to a family) and efficiency (the ratio of library events to matches) of waveform correlation. In this example using data from AAK, dramatic improvements resulted from using our Automated WCD compared with our Basic WCD—we doubled our percent of events that belong to a family and increased our efficiency manyfold. Our Automated WCD is continually being improved, and we expect to see additional improvements in both value and efficiency.

CONCLUSIONS AND RECOMMENDATIONS

Our research on using waveform correlation to process aftershock sequences strongly suggests that this is a useful technique for improving monitoring efficiency. Our most recent work in exploring how to automate the setup of the WCD has both improved performance and brought us one step closer to using WC in an operational system. However, much additional work needs to be done before that vision can be a reality.

We recommend that additional research explore the following tasks:

- Research how to optimize the selection of master waveforms to use in the library.
- Recognize a swarm has started, select station(s), and set up Master Waveform Library at station(s).
- Integrate waveform correlation results across a network: Our work to date has focused on using waveform correlation station-by-station. For an operational system, waveform correlation must be used for a network of stations. In further research we plan to explore how to combine the results from multiple stations.

ACKNOWLEDGEMENTS

Thank you to Richard Stead and David Yang at Los Alamos National Laboratory for sending us data on the Wenchuan earthquake.

REFERENCES

Harris, D. B. (1991). A waveform correlation method for identifying quarry explosions, *Bull. Seismol. Soc. Am.* 81, 2395–2418.

Thorbjarnardottir, B. S. and J. C. Pechmann (1987). Constraints on relative earthquake locations from cross-correlation of waveforms, *Bull. Seismol. Soc. Am.* 77, 1626–1634.

Withers, M., R. Aster, and C. Young (1999). An automated local and regional seismic event detection and location system using waveform correlation, *Bull. Seismol. Soc. Am.* 89, 657–669.

Wychecki-Vergara, S., H. Gray, and W. Woodware (2001). Statistical development in support of CTBT monitoring, DSWA01-98-C-0131.