Scalable Similarity Search in Seismology: A New Approach to Large-Scale Earthquake Detection

Karianne Bergen[1(✉)], Clara Yoon[2], and Gregory C. Beroza[2]

[1] Institute for Computational and Mathematical Engineering, Stanford University, Stanford, CA 94305, USA
kbergen@stanford.edu

[2] Department of Geophysics, Stanford University, Stanford, CA 94305, USA

Abstract. Extracting earthquake signals from continuous waveform data recorded by networks of seismic sensors is a critical and challenging task in seismology. Earthquakes occur infrequently in long-duration data and may produce weak signals, which are challenging to detect while limiting the number of false discoveries. Earthquake detection based on waveform similarity has demonstrated success in detecting weak signals from small events, but existing techniques either require prior knowledge of the event waveform or have poor scaling properties that limit use to small data sets. In this paper, we describe ongoing research into the use of similarity search for large-scale earthquake detection. We describe Fingerprint and Similarity Thresholding (FAST), a new earthquake detection method that leverages locality-sensitive hashing to enable waveform-similarity-based earthquake detection in long-duration continuous seismic data. We demonstrate the detection capability of FAST and compare different fingerprinting schemes by performing numerical experiments on test data, with an emphasis on false alarm reduction.

Keywords: Similarity search · Locality-sensitive hashing · Time series · Data mining · Earthquake detection · Template matching · Signal processing

1 Introduction

Seismology is an observational science that relies on data collected from seismic sensors to study and interpret processes within the earth. Earthquake detection, the use of signal processing to identify seismic signals in continuous ground motion measurements, is critical for enabling discoveries in the field. Modern seismic networks include hundreds to thousands of sensors, each recording data continuously. As the volume of available data grows, the seismology community is increasingly recognizing the need to adopt state-of-the-art algorithms and data-intensive computing techniques to process large seismic data sets.

There are a number of challenges and requirements for the earthquake detection problem. The events of interest, earthquakes, occur infrequently and their signals are short in duration (seconds to tens of seconds). Therefore, earthquake

© Springer International Publishing AG 2016
L. Amsaleg et al. (Eds.): SISAP 2016, LNCS 9939, pp. 301–308, 2016.
DOI: 10.1007/978-3-319-46759-7_23

detection requires processing months to years of data, most of which contains only background signals, including local, persistent noise sources. A practical earthquake detection system should be able to detect weak signals from small earthquakes while controlling the false alarm rate; a large number of false detections could easily overwhelm true events, so maintaining high precision is critical when processing large data sets. Small, low signal-to-noise events are hard to detect and can often only be confidently distinguished from noise by identifying coherent signals across an array of sensors. Sensor dropout and changes in sensor array configuration are not uncommon, so we focus on network-based detection approaches that detect independently on each channel as an initial step. This paper will focus on the single-channel detection problem.

The STA/LTA algorithm [1], widely used for general earthquake detection, identifies rapid increases in the signal energy to detect events with impulsive wave arrivals. This approach is attractive because it can be easily applied in near real-time to streaming data, but the simplicity of the detection statistic does not take advantage of the shape of the recorded waveforms.

Earthquake waveforms contain valuable information for detection; it has been widely observed that earthquakes originating at neighboring locations generate similar waveforms at a fixed sensor (Fig. 1). In recent years, seismologists have exploited waveform similarity, measured by the normalized cross-correlation, to detect small earthquakes with weak signals similar to those of known template events [6]. However, the performance of template matching is limited by the quality and availability of template waveforms from earthquake catalogs, which are known to be incomplete, especially for low magnitude events. We seek a general similarity-based earthquake detector that can identify similar earthquake waveforms without templates. Previous efforts toward that goal have proposed a brute-force blind search for similar waveforms [5], but the quadratic scaling of this approach makes it infeasible for large data sets.

In this paper we present on-going work to incorporate similarity search into a modern, scalable earthquake detection pipeline. We have introduced a new earthquake detection approach called Fingerprint and Similarity Thresholding (FAST) [8] to detect earthquakes by identifying similar waveforms in continuous

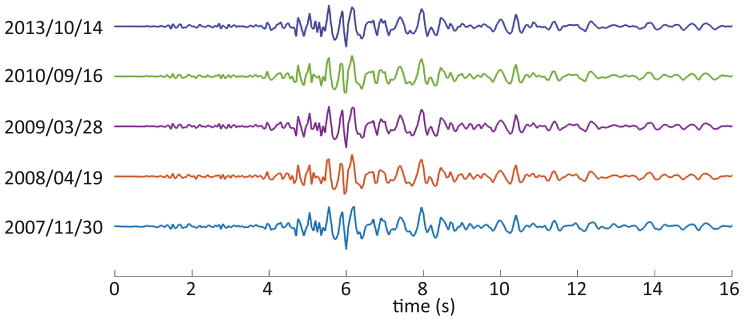


Fig. 1. Similar earthquake waveforms recorded during five distinct events over a period of years at a fixed sensor, station CCOB in Northern California

seismic data. FAST is modeled after scalable content-based audio identification systems [3]. Given continuous waveform data recorded by a single sensor, we extract a binary waveform fingerprint for each short-duration time interval. Then we perform an approximate similarity search using locality-sensitive hashing (LSH) to identify similar waveforms, which are labeled as candidate earthquakes. Below we describe our similarity-search-based approach for large-scale earthquake detection and discuss strategies to lower the false alarm rate and enable the detection of low signal-to-noise events.

2 Our Approach: FAST Earthquake Detector

The Fingerprint and Similarity Thresholding earthquake detection method identifies earthquakes using an efficient blind search for similar waveforms. The two key steps in the FAST detector are feature extraction and approximate similarity search. Feature extraction maps each short-duration waveform segment into a sparse binary fingerprint. The approximate similarity search, which employs locality-sensitive hashing [2] for computational efficiency, identifies similar pairs of fingerprints. Waveform segments corresponding to similar fingerprint pairs are classified as candidate earthquake signals.

2.1 Data

FAST operates on single-channel, continuous, high frequency (up to 100 Hz) data recorded by seismometers that measure ground motion at fixed locations. The data contain seismic signals embedded in background noise. We apply a 1–10 Hz bandpass filter and use a 10 s event window, corresponding to the predominant frequencies and duration of seismic waves for small local earthquakes.

2.2 Feature Extraction

Earthquake waveforms are searched using sparse binary waveform fingerprints. The feature extraction approach used in FAST is adapted from the Waveprint [3] method for audio fingerprinting. Audio fingerprinting provides a good starting point for the development of earthquake waveform fingerprints – there is structural similarity between the data and both applications require fingerprints that are robust to small variations and additive noise. The feature extraction process converts short-duration waveforms into sparse binary fingerprints (Fig. 2) and is described in the following steps.

1. **Spectrogram.** We convert the time series data to the spectrogram, a time-frequency representation computed with the short-time Fourier transform.
2. **Spectral Images.** We divide the spectrogram into short (10 second) overlapping segments, and resize spectral images to fixed dimensions: 32 frequency bins and 64 time bins. The spectral domain provides some shift invariance, unlike the time domain where waveforms must be precisely aligned; this allows a larger lag between adjacent intervals (1.0 vs. 0.05 s) and fewer fingerprints total, but the trade-off is reduced detection sensitivity.

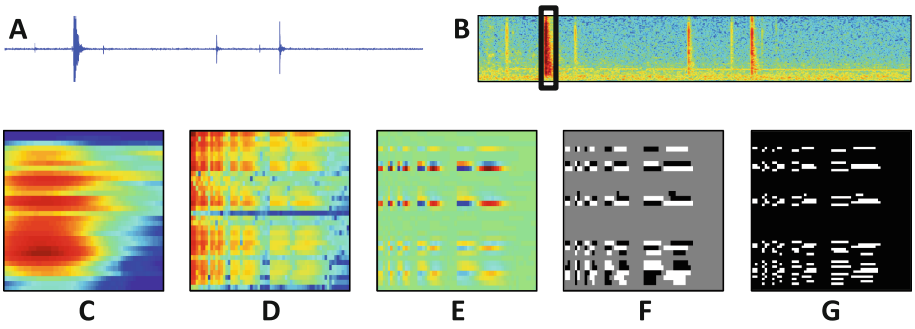


Fig. 2. Feature Extraction process in FAST: (A) continuous data, (B) spectrogram, (C) spectral image, (D) discrete Haar wavelet transform, (E) adjusted wavelet coefficients, (F) coefficient selection, (G) conversion to binary fingerprint

3. **Haar Wavelet Transform.** For each spectral image, we compute the two-dimensional discrete Haar wavelet transform.
4. **Coefficient Selection and Conversion to Binary.** We select the K *most anomalous* Haar coefficients (as described in Sect. 2.4) for each spectral image. K is typically selected in the range of 200–800 (out of 2048). For the selected coefficients, we retain only the sign value and set all other coefficients to zero. We convert the sign values to binary using two bits per coefficient, resulting in sparse binary fingerprints of dimension 4096 with K non-zeros.

2.3 Similarity Search

The computational efficiency of FAST comes from the use of locality-sensitive hashing [2] to perform a fast approximate similarity search. The Jaccard similarity coefficient quantifies the similarity between fingerprints. In the similarity search step, hash signatures are generated using MinHash [4] to preserve the Jaccard similarity, and LSH is used to identify fingerprints with similar signatures. The use of MinHash and LSH provides a significant improvement over the quadratic scaling of a brute-force all-to-all search. For instance, when applied to one week of continuous data, FAST has demonstrated a factor 140 speed-up over the brute-force search and detected 89 events compared with 24 events in the earthquake catalog (see [8] for details).

2.4 Haar Coefficient Selection

The effectiveness of similarity search is highly dependent upon the data representation. Fingerprints must be discriminative, that is similar waveforms map to similar fingerprints under the Jaccard metric. The imbalanced data set poses an additional challenge; the signals of interest, similar earthquake waveforms, appear infrequently in data dominated by background noise. In our template-free search, the potential for false detections is high because we search the full

seismic data record for similar pairs of waveforms to identify weak earthquake signals. Therefore we require fingerprints corresponding to background signals to be mutually dissimilar, even in the presence of persistent noise sources, to distinguish weak seismic signals while also limiting the number of false alarms.

The original feature extraction approach, following Waveprint, creates a compact representation using Haar wavelets by retaining the coefficients that are largest in magnitude. While this approach has been successfully applied in audio fingerprinting, when it is applied to seismic data the resulting fingerprints provide an inefficient representation; the largest magnitude coefficients often belong to a subset of frequently selected coefficients, while the majority of the coefficients are rarely selected. For instance, on a test data set with $K = 400$ selected coefficients, 16 % of the coefficients are "frequently selected" (i.e. active in at least 25 % of fingerprints) while 50 % are "rarely selected" (active in fewer than 1 % of fingerprints). This inefficiency impacts the performance of earthquake similarity search by increasing the average similarity between "background" fingerprints, thus making it more difficult to distinguish weak earthquake signals. Therefore we adjust our approach to select coefficients that are more discriminative with respect to background signals.

We select the Haar coefficients that the most discriminative or anomalous, rather than those that are largest in magnitude. To achieve this, we compute adjusted Haar coefficients by standardizing each coefficient based on its distribution across the full, background-dominated data set. We model the unknown coefficient distributions using simple statistics: with mean and standard deviation (Z-score), or with the median and median absolute deviation (MAD). These metrics allow us to choose coefficients that are not largest in magnitude, but farthest on the tails of the distribution. Empirically, this approach suppresses detections of persistent noise sources while maintaining high accuracy on earthquake signals (Fig. 3). We compare these fingerprinting schemes in Sect. 3.1.

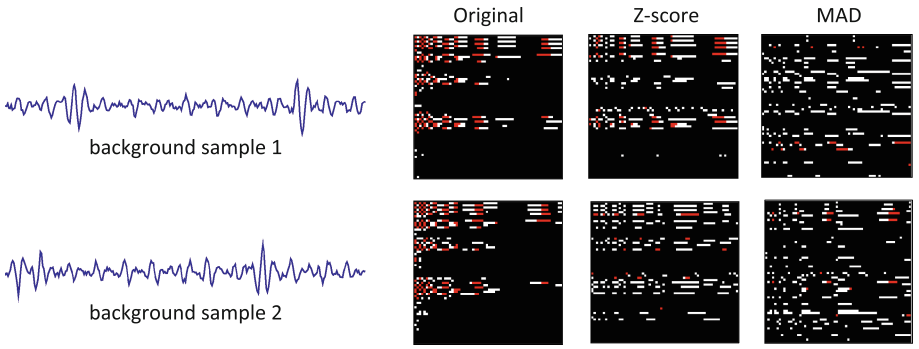


Fig. 3. Comparison of fingerprinting schemes applied to background noise. The Jaccard similarities between the fingerprints are: 0.266 (original), 0.117 (Z-score), and 0.040 (MAD).

3 Experiments

We compare the performance of the fingerprinting schemes described in Sect. 2.4 and demonstrate their accuracy for earthquake waveforms, then demonstrate the performance of FAST on a planted waveform test set in which earthquake waveforms are embedded in recorded background signals at known times and signal-to-noise ratio. All data used in the tests below were recorded at Northern California Seismic Network station CCOB, and sample earthquake waveforms were selected using the Northern California Earthquake Catalog.

3.1 Performance of Feature Extraction

We compare the three feature extraction schemes described in the previous section: (1) original, (2) Z-score-, and (3) MAD-adjusted fingerprints.

We test two criteria to measure the quality of fingerprints for our earthquake detection problem: fingerprint accuracy and baseline similarity. Accuracy is a measure of the quality of the fingerprints of earthquake waveforms for similarity-based detection under additive noise. Baseline similarity quantifies the similarity between background fingerprints in the presence of persistent noise to estimate false detection rates.

To assess accuracy, we compare the fingerprints of clean earthquake waveforms to low signal-to-noise versions of the same waveform embedded in noise:

$$\text{accuracy}(i,j) = \text{jaccard}\left(\mathcal{F}_{\mathcal{P}}(x^{(i)}),\, \mathcal{F}_{\mathcal{P}}(\alpha x^{(i)} + n^{(j)})\right), \tag{1}$$

where $\mathcal{F}_{\mathcal{P}}$ is the feature extraction operation, $x^{(i)}$ is the i-th earthquake waveform, $n^{(j)}$ is the j-th background waveform, and α is a scaling factor to control the signal-to-noise ratio (SNR). We use waveforms from 300 known earthquakes and embed each one in 10 noise segments at a low SNR ranging from 1.0 to 5.0. To test the robustness of the fingerprints, the signals were bandpass filtered to 1–10 Hz and include persistent noise in the 1.5–3.5 Hz range. We directly compute the Jaccard similarity between the clean and noisy fingerprints and report the median for each feature extraction scheme in Table 1. The MAD-adjusted fingerprints consistently have the highest accuracy.

Table 1. Median Jaccard similarity of clean and low-SNR earthquake waveforms

SNR	Fingerprint accuracy		
	Original	Z-score	MAD
1.0	0.3093	0.3629	0.4760
2.0	0.5123	0.6736	0.7279
4.0	0.7354	0.8561	0.8735

The baseline similarity distribution is estimated from the Jaccard similarities for 5000 pairs of background fingerprints:

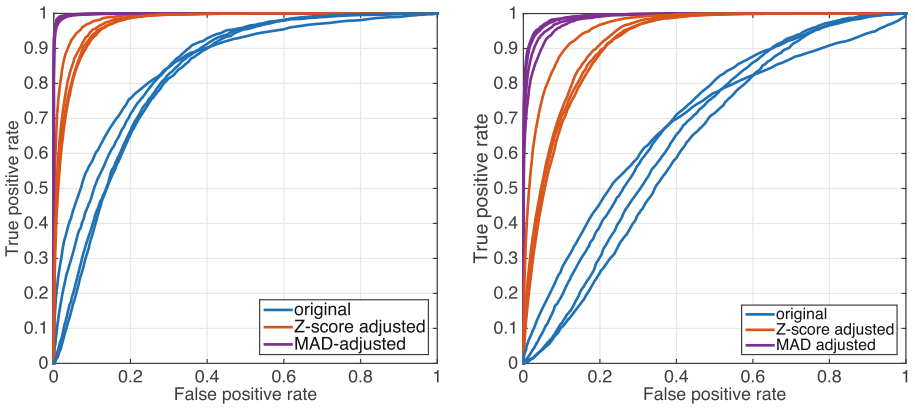
$$\text{baseline}(k,\ell) = \text{jaccard}\left(\mathcal{F}_{\mathcal{P}}(n^{(k)}),\ \mathcal{F}_{\mathcal{P}}(n^{(\ell)})\right). \tag{2}$$

The similarity between background fingerprints is substantially lower for MAD- and Z-score adjusted fingerprints than for the original top magnitude fingerprints, with median Jaccard similarities of 0.047, 0.071, and 0.185, respectively.

'In order to maintain high overall precision in an imbalanced data set, we require both high accuracy for fingerprints and low baseline similarity to limit false detections. We characterize the trade-off between false detections and missed detections, specifically for the case of identifying low SNR earthquakes similar to clear earthquake waveforms, using in a ROC curve (Fig. 4a). For a given Jaccard similarity threshold, the *true positive rate* is defined as the rate at which the accuracy exceeds this threshold, and the *false positive rate* is the rate at which the baseline similarity exceeds the same threshold. We also consider the more challenging and relevant case in which we seek to identify pairs of similar low SNR earthquake waveforms, i.e. both instances of the waveform include additive noise in a modified accuracy formula (Fig. 4b).

3.2 Detection Performance

To have a clear ground truth for measuring detection performance, we inject real earthquake waveforms into a dataset consisting of 16 hours of recorded background signal. Twelve pairs of known event waveforms are embedded in the background at low SNR. We report the results for MAD-adjusted fingerprints with $K = 400$ non-zeros, and 100 hash tables with 4 hash functions per table in the LSH search. The detection statistic is the fraction of hash tables in which a fingerprint appears in the same hash bucket as its nearest neighbor. FAST successfully identifies all 24 low SNR events with only 4 false detections (85.71 % precision). FAST has shown promising initial results on real earthquake sequence



(a) Clean-to-noisy earthquake waveforms (b) Noisy-to-noisy earthquake waveforms

Fig. 4. Trade-off between detection rate for weak signals (SNR 1.0) and false detections. Multiple lines represent results for several different values of K.

data, detecting previously unknown events with a manageable number of false detections in months of continuous data.

4 Discussion

In this paper, we present an application of approximate similarity search with LSH to the problem of earthquake detection in continuous seismic data. This work represents a new direction for waveform-similarity-based earthquake detection that does not require prior knowledge of event waveforms and has sufficient computational efficiency to allow for application to long-duration data that would not be feasible using a brute-force search. Our initial experiments with FAST demonstrate that this approach can successfully detect previously unknown small earthquakes using blind similarity search. Furthermore, we have demonstrated that modifications to audio fingerprinting methods based on the empirical data distribution can improve accuracy on imbalanced data sets, which contain relatively few pairs of moderate-to-high similarity. Scalable similarity search has the potential to impact both the study of earthquakes and earth and environmental monitoring more broadly. Imbalanced data sets appear in many of these applications, such as acoustic recordings used for mining bioacoustic soundscapes in ecological studies [7], and we believe the techniques developed for FAST can be applied in these domains.

Acknowledgments. This research was supported by NSF grant EAR-1551462 and by the Southern California Earthquake Center (contribution no. 6325). Waveform data, metadata, or data products for this study were accessed through the Northern California Earthquake Data Center, doi:10.7932/NCEDC. We thank Ossian O'Reilly for his assistance with the hashing techniques used in this work.

References

1. Allen, R.: Automatic phase pickers: their present use and future prospects. Bull. Seismol. Soc. Am. **72**(6B), S225–S242 (1982)
2. Andoni, A., Indyk, P.: Near-optimal hashing algorithms for approximate nearest neighbor in high dimensions. Commun. ACM **51**(1), 117–122 (2008)
3. Baluja, S., Covell, M.: Waveprint: efficient wavelet-based audio fingerprinting. Pattern Recogn. **41**(11), 3467–3480 (2008)
4. Broder, A.Z., Charikar, M., Frieze, A.M., Mitzenmacher, M.: Min-wise independent permutations. J. Comput. Syst. Sci. **60**(3), 630–659 (2000)
5. Brown, J.R., Beroza, G.C., Shelly, D.R.: An autocorrelation method to detect low frequency earthquakes within tremor. Geophys. Res. Lett. **35**(16), L16305 (2008)
6. Gibbons, S.J., Ringdal, F.: The detection of low magnitude seismic events using array-based waveform correlation. Geophys. J. Int. **165**(1), 149–166 (2006)
7. Servick, K.: Eavesdropping on ecosystems. Science **343**(6173), 834–837 (2014)
8. Yoon, C.E., O'Reilly, O., Bergen, K.J., Beroza, G.C.: Earthquake detection through computationally efficient similarity search. Sci. Adv. **1**(11) (2015)