PageRank for Earthquakes

by Ana C. Aguiar and Gregory C. Beroza

INTRODUCTION

Using Hi-Net data in southwest Japan, Obara (2002) discovered a new type of seismic source that he called nonvolcanic tremor due to its similarity to volcanic tremor, but distinct origin. Tremor is characterized by weak signals, typically of deep (>30 km) origin, that are most prominent in the 1–10 Hz frequency band. At about the same time, slow-slip events and tectonic tremor were found to be strongly correlated indicating that tremor accompanied these events (Obara, 2002; Rogers and Dragert, 2003). Tectonic tremor is challenging to locate because it is an extended and persistent signal, and lacks the clear *P*- and *S*-wave arrivals that are typical of earthquakes. Despite this, there is clear modulation of tremor amplitudes that can be tracked between stations and used to locate the tremor source. Several envelope techniques (Obara, 2002; McCausland *et al.*, 2005; Wech and Creager, 2008; Aguiar *et al.*, 2009) were developed for this purpose.

Low-frequency earthquakes (LFEs) were discovered by Katsumata and Kamaya (2003) and have since proven to be critical to understanding the genesis of tectonic tremor. Shelly *et al.* (2006) showed that LFEs correlate strongly with times of tremor and that they occur on the plate interface. Shelly *et al.* (2007a) showed that tremor consists of swarms of LFEs, and by inference occurred on the plate interface as well. Follow-up work (Ide, Shelly, Beroza, 2007) demonstrated that LFEs per tremor occurred as shear slip and hence were part of a family of slow earthquakes (Ide, Beroza, *et al.*, 2007).

Initially, LFEs were found through visual inspection by enterprising network analysts in Japan (Beroza and Ide, 2011). LFEs have been identified as periods of high-amplitude arrivals by the same visual inspection technique (Shelly, 2009). Because LFEs have proven so useful for studying tremor, other methods have been developed to identify them.

First, LFEs from the Japan Meteorological Agency (JMA) catalog were used as templates to find other LFEs by cross correlating the signal of the LFE with tremor (Shelly *et al.*, 2007a,b). For the cases where no previous templates exist, Brown *et al.* (2008) developed an autocorrelation method to find the LFEs. This method considered each small window as a possible detection by cross correlating each window with all the others (Brown *et al.*, 2008) and searching for repeats. To date, LFE detection methods do not fully exploit the fact that LFEs repeat, with similar waveforms, not once, but multiple times.

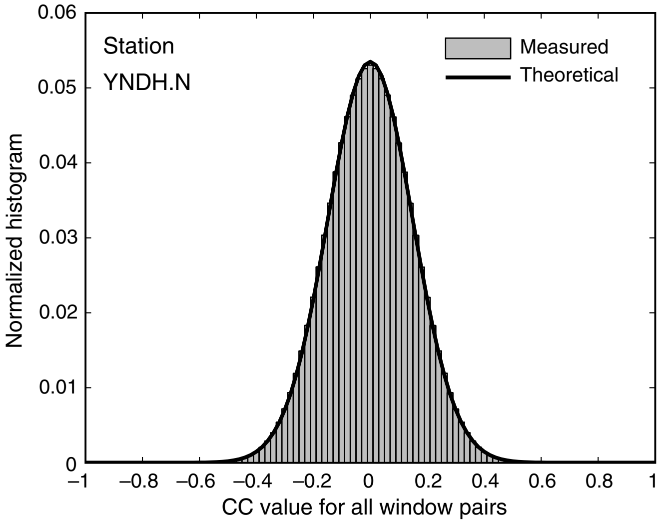
In this study, we take advantage of the similarity of LFEs within tremor by developing a test of significance that explicitly accounts for short-time correlations in the data, and the likelihood that the sources repeat more than once. To do this, we use the structure of how similar windowed waveforms are linked with one another. A complicating factor is that these links do not form closed loops that satisfy closure, but have links that may be incomplete. Despite this lack of closure, we would like to take full advantage of the structure of similar seismograms. Our problem maps very closely to one that has been solved previously. One of the initial web page searching algorithms implemented by Google used the links between pages to calculate the quality of a document. This data mining algorithm is known as PageRank and was developed by Page *et al.* (1999). If we relate this to our problem, but substitute candidate windows in place of web pages, and we know that these windows are linked to one another, we can use the PageRank algorithm to rank them. In applying PageRank to tremor data, we find the ranking of a specific window and show statistically how windows rank relative to each other. This tells us which windows are most likely potential LFEs, and can be used to construct templates through waveform stacking. These stacks provide high signal-to-noise ratio (SNR) templates that can be used to find other LFEs.

We use the April 2006 Shikoku, Japan, tremor episode to test our method due to the wealth of available information concerning tremor in this area (Obara, 2002; Shelly *et al.*, 2006, 2007a; Ide, 2010) and this specific tremor episode has been analyzed in great detail (Shelly *et al.*, 2007b). By applying the PageRank approach we create robust LFE templates that match known LFEs from the JMA catalog. Using these templates we find detections in data for the April episode with similar results to Shelly *et al.* (2007b) for the same time periods. We also find detections for weaker segments of tremor that were not previously reported. Our approach suggests a new approach to differentiate tremor signals from noise for sparse data sets using the fact that PageRank behavior is qualitatively different for tremor versus noise.

MULTIPLICITY OF REPEATS

To find LFEs within the tremor data, we begin with the autocorrelation method of Brown *et al.* (2008), which detects potential LFEs in a pair-wise manner. This method finds the signals by autocorrelating each window with all other windows in time during a tremor segment of interest. We divide the tremor data in 10 s windows lagged by 0.08 s or 2 samples for data with 25 samples per second. This creates a population of 44,900 windows for 1 hour of data. The population of values for the correlation coefficients (CC) between these windows closely follows a normal distribution (Fig. 1). We

 doi: 10.1785/0220130162



▲ **Figure 1.** Distribution of correlation coefficient (CC) values for all window pairs of 1 hour of tremor data on station YNDH.N. The black line represents the theoretical normal distribution calculated using $\sigma = 1.253 \times \text{MAD}$. Here, MAD is computed from all window pairs.

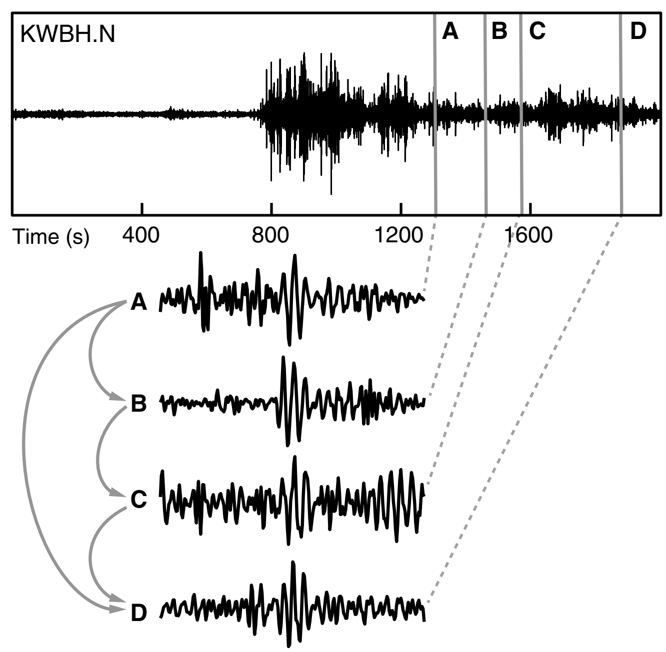
have applied the Fisher transformation to the CC values, in which

$$z = \frac{1}{2}[\ln(1 + \text{CC}) - \ln(1 - \text{CC})]. \tag{1}$$

This transformation accounts for the fact that the distributed values cut off at $\text{CC} = \pm 1$. With this, we can assume normal statistics for the distribution of the CC values. We find that in our case the transformation made no significant difference in the distribution, due to the small values of CC (Fisher, 1973). For this reason, we base our detection threshold on a normal distribution with zero mean and focus on the large positive values to declare a positive detection. For a normal distribution, $\sigma = 1.253 \times \text{meanabsolutedeviation (MAD)}$. MAD is the mean of the absolute deviations of a set of data about the data's mean and is a measure that is chosen to be insensitive to outliers. Our objective is to find the positive outliers in the data, which represent the repeating signals within the tremor. Given that the MAD is not affected by outliers it will not bias our estimate of σ, so we can use this as a measure to establish a threshold of detection. By assuming zero mean (Fig. 1), we can calculate the MAD directly from the population of CC values and estimate σ from this. Here we have chosen to use the mean instead of the median which has been preferred in previous analyses (Shelly *et al.*, 2006; Brown *et al.*, 2008). We found that for a very large set of window pairs the computational costs of calculating the median are substantial whereas the mean is easily computed. With this, we perform the autocorrelation to find detection pairs for each of the stations analyzed with a low positive threshold of 3σ, which corresponds to a two-sided significance level of 99.7%. This level, which we use to define a positive correlation, is a trade-off between a higher threshold, which will provide more confident matches at the cost of fewer positives for our low-SNR data, and a lower threshold, which will provide more positives, but with less confidence in the reliability of individual matches. The noise level, and therefore the statistical behavior of the autocorrelations, varies between stations so we perform the analysis one station at a time.

Finding detection pairs within the tremor data does not ensure closure because LFEs repeat multiple times during a single tremor episode (Shelly *et al.*, 2007a) and the SNR is low. Figure 2 shows this lack of closure schematically. Window A correlates significantly with Window B and Window D. Later in time, Window B matches with Window C, but Window A does not match with Window C, and Window B does not match with Window D. Such a lack of closure is inevitable with low-SNR signals, and complicates detection statistics, yet we have to make the best of the information we have. That is, we want to use all the links, and the complex hierarchical relationships they express, to identify the windows most likely to represent LFE signals.

We use a tool developed for data mining to address this problem. Specifically we apply the PageRank algorithm to the detection pairs from the autocorrelation process to determine which windows have the most number of links. We then calculate a ranking for each window and if the probability is high, this suggests a robust LFE detection for that window within the data.



▲ **Figure 2.** Window pair links: A, B, C, and D are windows that were found to be a match with another window, given the threshold selected. This figure shows how all these pairs are linked together. Some windows are linked directly: A → B and B → C. Other windows are linked indirectly: A is linked indirectly to C (because B → C), and B is indirectly linked to D (because C → D).

THE PAGERANK APPROACH

The PageRank method calculates the probability of importance of each web page in a set of web pages, assembled in a vector. These probabilities are based on the number of links each web page has, assigning a high probability to web pages with high number of links and low probability to pages with low number of links. Specifically, each of the elements of this vector is the probability that a random surfer will visit a particular page on the web (Page *et al.*, 1999; Moler, 2011). Like a Markov Chain, PageRank is a random, iterative method where the probability at one stage in the iterations is computed from the previous stage, with an initial condition of equal probabilities for all pages. For our problem of seismic data, we form a connectivity matrix **G** from the detection pairs found during the autocorrelation that contains all the information on links between all windows. If we find a match between window i and window j in time, then $g_{ij} = 1$, otherwise $g_{ij} = 0$. Now, if we consider p to be the probability of the random walk following a specific link to a window, and the probability of an arbitrary window to be chosen as $1 - p$, then the probability that a certain random window is chosen will be $\delta = (1 - p)/n$, in which n is the total number of windows. With this, we form a matrix **A** that scales the **G** matrix by the sum of its columns:

$$a_{ij} = \begin{cases} pg_{ij}/c_j + \delta: & c_j \neq 0 \\ 1/n: & c_j = 0 \end{cases}, \quad (2)$$

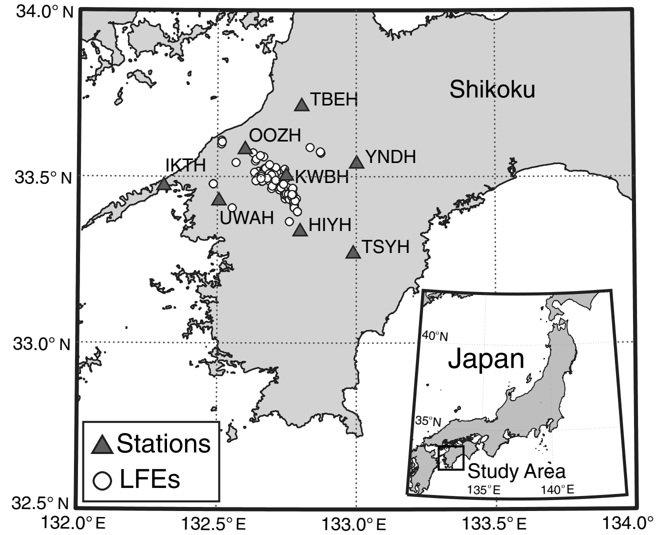
in which $c_j = \sum_i g_{ij}$ and **A** is the transition probability matrix (Moler, 2011). We then solve equation (3) iteratively to find the PageRank:

$$\mathbf{x} = \mathbf{A}\mathbf{x}. \quad (3)$$

This equation has a unique, nonzero solution if a scaling factor is chosen such that $\sum_i x_i = 1$. Given this condition, **x** is the steady state vector of **A** and therefore the PageRank (Moler, 2011).

As a starting point, we assign equal probabilities to each window of $1/n$ and then iterate equation (3) until **x** converges to a desired tolerance. Here, we have selected a standard value of $p = 0.85$ and a tolerance of 0.01 to calculate the PageRank for each window in the tremor analysis. As we iterate, we calculate the 1-norm between the probabilities of the previous and current iteration to estimate the tolerance. If it is below $0.01/n$, we stop the iteration and use that **x** for PageRank.

To test our method, we use data from eight stations in Shikoku Japan (Fig. 3) during the April 2006 tremor event. This group of stations was selected due to the locations of a large number of well-studied LFEs that have occurred within this area (Shelly *et al.*, 2007b). We analyzed 1 hour of data at 25 samples per second from 16 April 2006, where a large number of LFEs have been previously detected using templates (Shelly *et al.*, 2007b). We chose a time period where the tremor signal is small at the beginning and increases in amplitude toward the end (Fig. 4) to understand differences in the behavior of PageRank. Using these data, we first calculated the autocorrelations using the population of 44,900 windows created from the 10 s windows lagged by two samples. We then applied the PageRank approach to the detection pairs to find the probabilities for each window. Figure 4 shows the PageRank values for each window in one hour of Japan data for several stations.

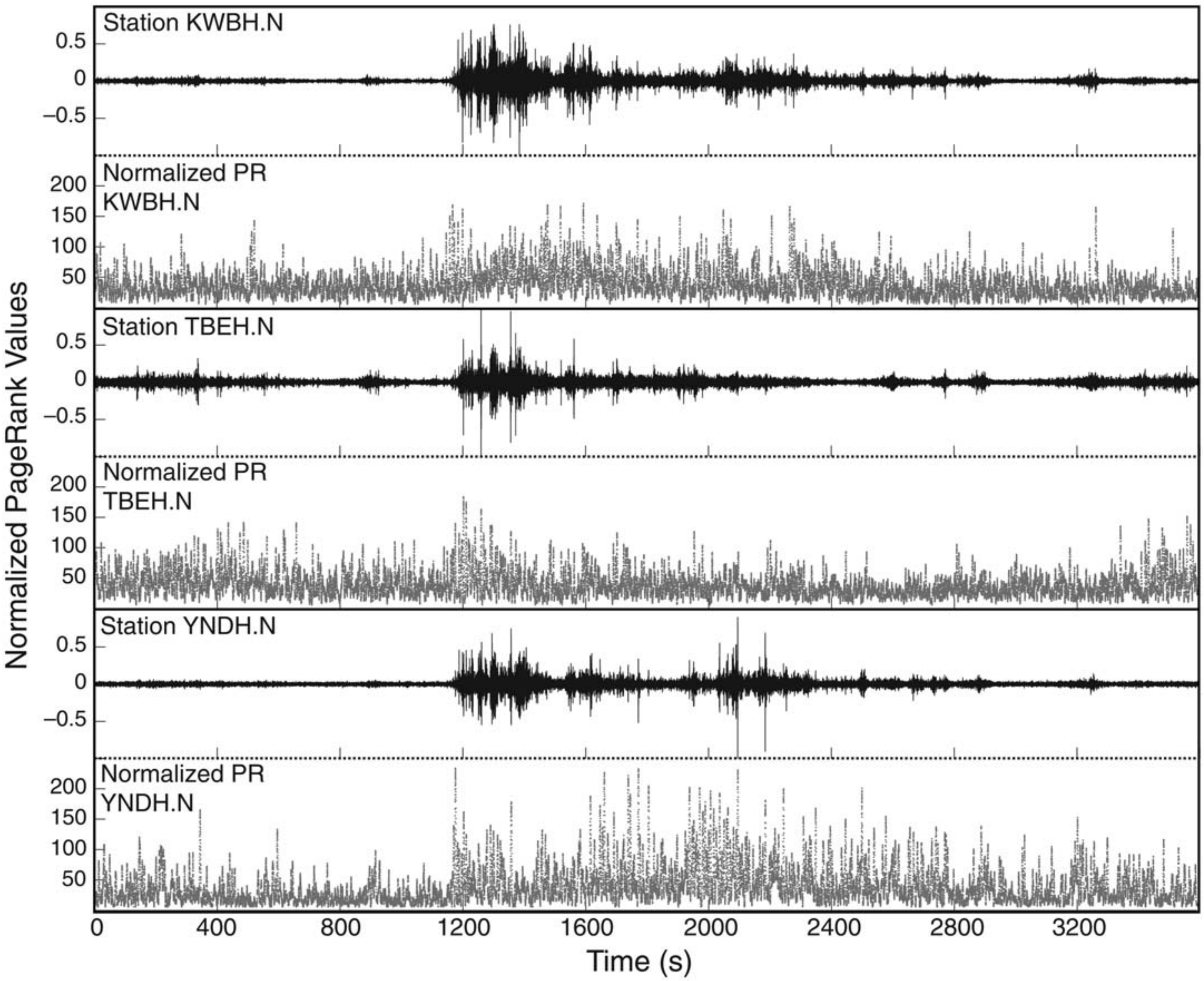


▲ **Figure 3.** Map of Shikoku, Japan, with the location of stations (dark gray triangles) used in our analysis. Circles represent LFE locations from Shelly *et al.* (2007b) during the tremor episode of 16 April 2006. The inset map shows the location of the study area in Japan.

Once we have computed the PageRank for each window, we know which windows have the highest probabilities of being repeating signals, and are the most likely to be LFE waveforms. A large PageRank value here signifies a window with a large number of links to other windows, both direct and indirect. Figure 4 shows that the PageRank values get higher as the tremor amplitude increases for all three stations shown. It is also noteworthy that there are high values toward the beginning where tremor amplitude is smaller.

We use the windows with the highest PageRank values to create template signals by stacking the matched windows. To exploit the multiplicity of the LFEs, we use both the direct and indirect matches of the high-ranked windows (Fig. 4) to create the stack for the template LFE. Figure 5 shows a stack with all the windows that were found to be a match to a window with a high number of matches for station YNDH.N during the one hour of tremor data analyzed. On the left, the stacked signal (Fig. 5a) is formed by windows with direct links only. On the right (Fig. 5b), the stacked signal includes windows from both direct and indirect links to the window with high PageRank.

We form the template from both the windows that directly match the main window, and also all the windows that matched each one of those initial matches. These are the windows with indirect links to the main window. The advantage of this process can be seen in Figure 6, which shows a comparison between the first, second and third level stacks for several stations used in the analysis. The signal of the stacked waveform improves

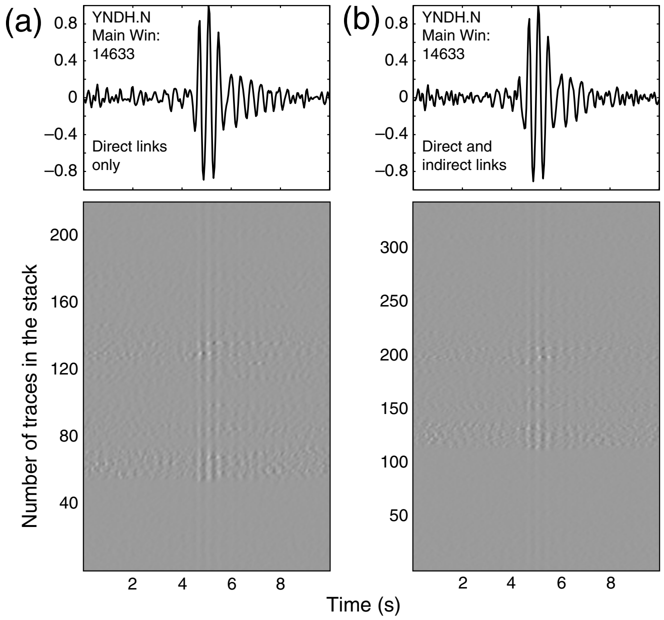


▲ **Figure 4.** One hour of tremor data (black) during the tremor event in Shikoku, Japan, on 16 April 2006 for stations KWBH, TBEH, and YNDH associated to its PageRank values (gray). The PageRank values are normalized by the total number of windows analyzed.

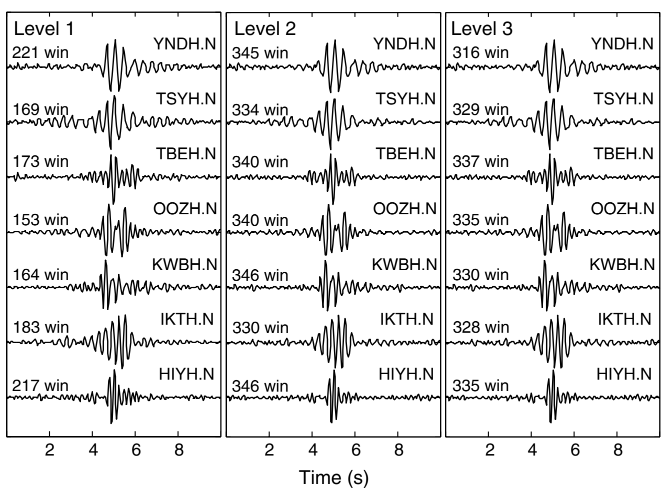
significantly once a second level of detections is added to the stack (windows with indirect links). The difference is apparent from level 1 to level 2 (Fig. 6), where each signal has less noise around the main peak of the largest amplitude arrival.

Here, each level of stack was created with a different number of windows. In the case of level 1, the lowest number of windows was used because these stacks only include the windows with direct links to the highest ranked window. Each window from the direct links has its own set of links. These would be the first level of windows with indirect links to the main window. Level 2 has a larger number of stacks than level 1 because it also includes these windows, the first level of windows with indirect links. Adding these windows to each stack improves the signal for each template (Fig. 6).

We can add another level of links by stacking the windows with links to the first level of windows with indirect links. These stacks are shown in Figure 6 as level 3. We can see that adding one more level of windows does not change the stack significantly and the number of signals used is very similar to level 2. The reason for this is that the windows added at this step are mostly windows that were already present as links to other windows. For that reason they do not contribute independent information to the stack. We note here that in some cases the number of windows used for level 3 is slightly lower than in level 2. This is caused by the way we have defined the near repeat window elimination in the processing. As we scan the list of windows, we eliminate the near repeats by finding the window that was paired with the best CC value. We do this in ascending numerical order for a 3 s time span. If the window with the highest CC was not the first on that group then this pushes the selected window number forward. These might cause the next group of windows to be considered now as a near repeat if the window number of the first of the group is now less than 3 s than the previously selected one. Adding



▲ **Figure 5.** Stacked signal (top) formed by summing all the windows (a) using only windows with direct links (221 traces) and (b) using windows with direct and indirect links (345 traces) that were found as a match to the window with the highest PageRank. The grayscale plot (bottom) of each one of the windows forming the stack.

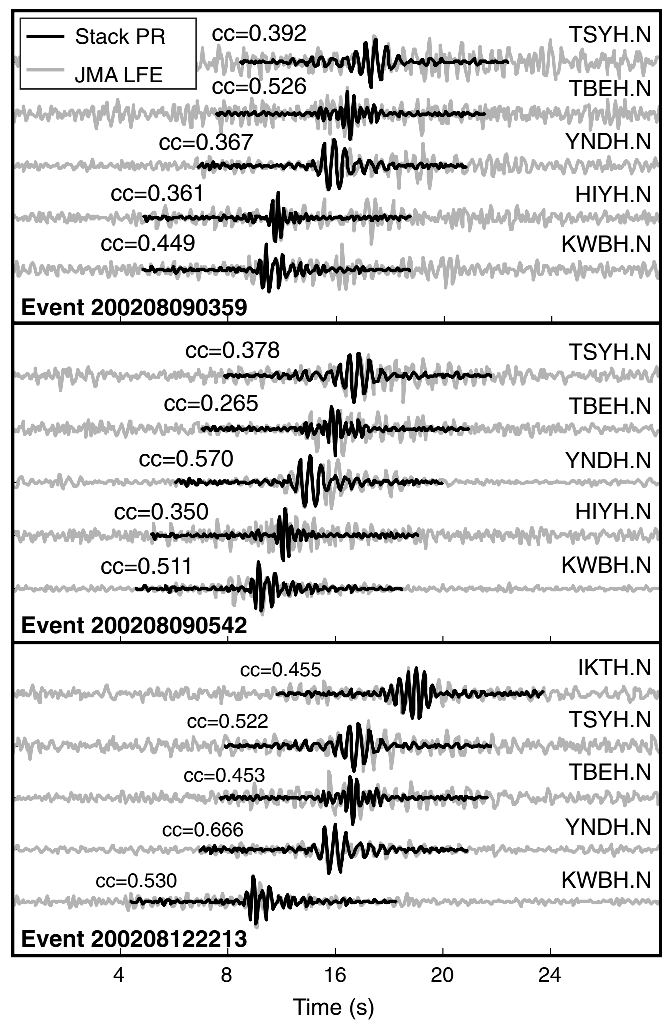


▲ **Figure 6.** Plot of stacked signals formed by windows of three different levels of links. Level 1 shows stacks created using only the windows with direct links to the window with highest PageRank. Level 2 shows stacks created with windows with direct links and one level of indirect windows (windows matched to the matches of the main window). Level 3 shows the stacks created with direct windows and two levels of indirect windows. The number on the left of each plot shows the number of traces that were used to create each stack.

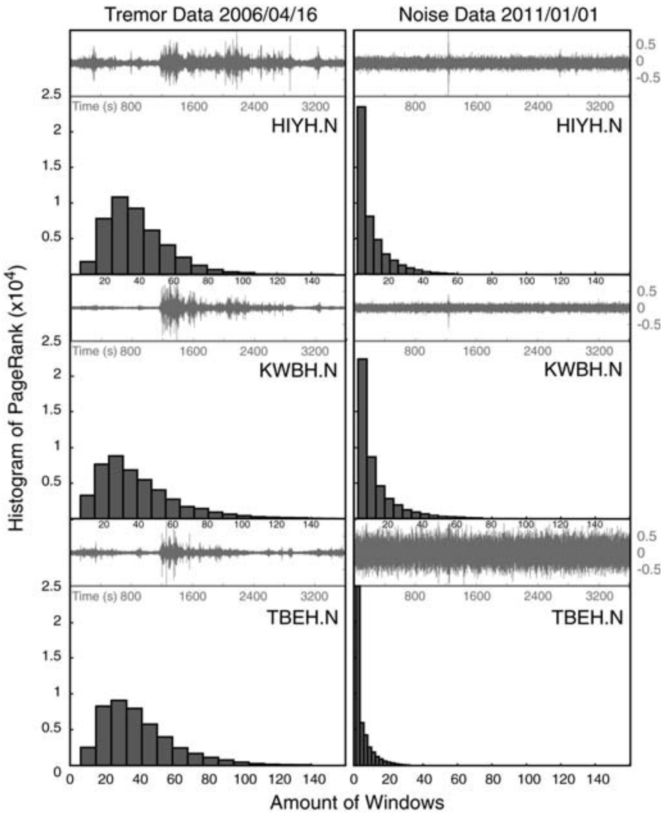
one more level of windows only finds more similar windows to the signals used in level 2 and some of these might fill some small gaps in, causing a few more near repeats to be eliminated. Given the similarity of the stacks from level 2 and level 3 (Fig. 6), this suggests that it is only necessary to create the stacks using one level of windows with indirect links to get an optimum result for creating a template waveform.

We validate the templates created with these stacks by comparing the stacked signals to known LFEs from the JMA catalog. We selected various events from the catalog that locate in the same general area as the LFEs reported for the April 2006 sequence (Shelly *et al.*, 2007b). Figure 7 shows the waveform comparison between the LFEs from the JMA catalog and the templates created in this study for the stations available in the catalog. The stacks created for the five stations, shown in Figure 7, match the three different events from the catalog and preserve the move-out across the network.

A significant advantage to the PageRank approach is that it can also be used to distinguish between tremor and noise. We



▲ **Figure 7.** Three different LFEs from the JMA catalog (gray) compared with one template stack (black) created using the PageRank approach from an hour of Shikoku tremor data from 16 April 2006.



▲ **Figure 8.** Histograms of normalized PageRank values for tremor data (left) and noise data (right) for three different stations used in the analysis. Noise data shows large numbers of low PageRank values, whereas the tremor data have higher PageRank values.

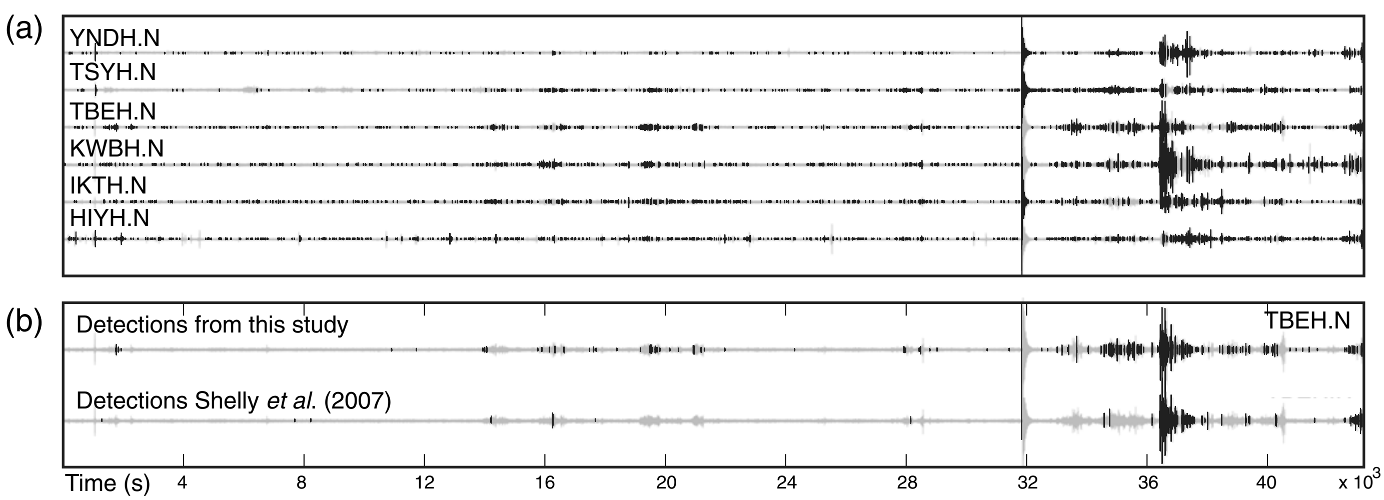
compared an hour of tremor data from the 16 April 2006 episode to an hour of noise data for the same stations in Shikoku. Figure 8 shows the differences between the normalized PageRank histograms for three different stations used in this analysis. These histograms are significantly different between tremor and noise because tremor has many more windows with high PageRank.

If the data has a lower SNR, then this characteristic of PageRank could be used to help distinguish tremor from noise. Because the differences show up in single components at different stations independently, it should also help to distinguish tremor from noise for sparse data.

APPLICATION TO CONTINUOUS DATA: SOUTHWEST JAPAN

Having found robust signals to use as LFE templates, we applied these to search through continuous data during the Shikoku event in April 2006. We picked data where it is clear that the strong tremor episode is getting started, so we can observe small tremor bursts but also larger amplitude, more significant bursts (Fig. 9) to test the ability of our detector to find LFEs within lower SNR tremor data. We cross correlate the templates with 10 s windows and move the window every two samples through the data to find matches. We perform this analysis one station at a time using a low threshold (3σ) and later compare the results between stations to distinguish between true and false detections.

Figure 9 shows the second half of 16 April 2006 for Shikoku Hi-Net data. This data set shows small tremor bursts between 12,000 and 24,000 s and larger amplitude tremor bursts toward the end of the day, around 32,000 s. Here we used only stations were it was possible to compare our LFE template to LFE picks present in the JMA catalog (Fig. 7). To associate detections found for each station (Fig. 9a), we compare all stations within a 2 s window and declare a positive detection if three or more of the stations show a detection within those 2 s. Using this simple association algorithm we find very similar results to Shelly *et al.* (2007b), particularly for the stronger tremor burst toward the end of 16 April 2006. We also find a large number of detections that were missed previ-



▲ **Figure 9.** Half day of data during 16 April 2006 of several stations in Shikoku showing the beginning of the tremor episode. (a) Detections (black) found individually for each station using the stack created with the PageRank approach for six of the stations in the analysis. (b) Detections found on at least three of the stations within a 2 s window (top), compared to Shelly *et al.* (2007b) detections (bottom) using template matching.

ously, during smaller tremor bursts (Fig. 9b) between 12,000 and 24,000 s and between 32,000 and 36,000 s.

DISCUSSION AND CONCLUSIONS

We use the PageRank approach to detect LFEs during the April 2006 tremor episode in Shikoku Japan. We selected this data set to facilitate the comparison of results with Shelly *et al.* (2007b) to test our method's ability to detect LFEs within data with both high and low SNR. Analyzing one station at a time, and applying low initial detection thresholds, we created robust LFE templates that match real LFE signals from the JMA catalog for this tremor episode without the use of any previous knowledge of event times. These templates include windows with both direct and indirect links to the highest PageRank window, which improves the SNR of the template created for each station. These templates will facilitate the detection of other events within the data.

We cross correlated the stacks we developed using the PageRank approach to find other LFEs within the April 2006 tremor episode in Shikoku, Japan. We find that our detections are comparable to detections found by Shelly *et al.* (2007b) where the tremor burst has large amplitudes compared to the rest of the time. We also find a number of detections within smaller tremor bursts that were previously missed. This suggests that the PageRank approach is a good tool for finding LFEs within lower SNR data.

We have also found that the PageRank distribution differs between tremor and noise. By looking at histograms of PageRank values we can differentiate between tremor and noise because unlike noise, tremor data has large numbers of high PageRank values due to the significant links between windows within a short period of time. This could prove to be a useful tool for automated tremor detection and for detecting tremor where it has not been detected previously, or where it has been found but the signal is not as prominent as it is on Hi-Net data. Finally, although we have applied it to detect LFEs within tremor, the PageRank approach may be useful for other situations such as swarms or aftershock sequences, for which many similar waveforms may be present.

ACKNOWLEDGMENTS

We want to give a special acknowledgment to Andrew Bradley for his help and direction on implementing the PageRank algorithm. We also want to thank Justin Brown and David Shelly for providing helpful discussion. All data for this work were obtained from the NIED Hi-Net data server, and we utilized the Stanford Center for Computational Earth and Environmental Science for most processing. This research was supported by the Southern California Earthquake Center (SCEC) and U.S. Geological Survey (USGS) Grant G10AP000160001. SCEC is funded by National Science Foundation Cooperative Agreement EAR-0529922 and USGS Cooperative Agreement 07HQAG0008. The SCEC Contribution Number for this paper is 1803.

REFERENCES

Aguiar, A. C., T. I. Melbourne, and C. W. Scrivner (2009). Moment release rate of Cascadia tremor constrained by GPS, *J. Geophys. Res.* **114,** no. B7, B00A05, doi: 10.1029/2008JB005909.

Beroza, G. C., and S. Ide (2011). Slow earthquakes and nonvolcanic tremor, *Ann. Rev. Earth Planet. Sci.* **39,** no. 1, 271–296.

Brown, J. R., G. C. Beroza, and D. R. Shelly (2008). An autocorrelation method to detect low frequency earthquakes within tremor, *Geophys. Res. Lett.* **35,** no. 16, L16305, doi: 10.1029/2008GL034560.

Fisher, R. A. (1973). *Statistical Methods for Research Workers*, 14th Edition, Revised and Enlarged Edition, Hafner Publishing Company, New York.

Ide, S. (2010). Striations, duration, migration and tidal response in deep tremor, *Nature* **466,** no. 7304, 356–359.

Ide, S., G. C. Beroza, D. R. Shelly, and T. Uchide (2007). A scaling law for slow earthquakes, *Nature* **447,** no. 7140, 76–79, doi: 10.1038/nature05780.

Ide, S., D. R. Shelly, and G. C. Beroza (2007). Mechanism of deep low frequency earthquakes: Further evidence that deep non-volcanic tremor is generated by shear slip on the plate interface, *Geophys. Res. Lett.* **34,** no. 3, L03308, doi: 10.1029/2006GL028890.

Katsumata, A., and N. Kamaya (2003). Low-frequency continuous tremor around the Moho discontinuity away from volcanoes in the southwest Japan, *Geophys. Res. Lett.* **30,** no. 1, 1020.

McCausland, W., S. Malone, and D. Johnson (2005). Temporal and spatial occurrence of deep non-volcanic tremor: From Washington to northern California, *Geophys. Res. Lett.* **32,** no. 24, L24311, doi: 10.1029/2005GL024349.

Moler, C. (2011). Experiments with MATLAB, Edited, MathWorks, Inc., www.mathworks.com/moler/exm/chapters.html (last accessed November 2013).

Obara, K. (2002). Nonvolcanic deep tremor associated with subduction in Southwest Japan, *Science* **296,** no. 5573, 1679–1681.

Page, L., S. Brin, R. Motwani, and T. Winograd (1999). The PageRank citation ranking: Bringing order to the Web, *Technical Rept.*, Stanford InfoLab, 5 pp.

Rogers, G., and H. Dragert (2003). Episodic tremor and slip: The chatter of slow earthquakes, *Science* **300,** 1942–1944.

Shelly, D. R. (2009). Possible deep fault slip preceding the 2004 Parkfield earthquake, inferred from detailed observations of tectonic tremor, *Geophys. Res. Lett.* **36,** no. 17, L17318, doi: 10.1029/2009GL039589.

Shelly, D. R., G. C. Beroza, and S. Ide (2007a). Non-volcanic tremor and low-frequency earthquake swarms, *Nature* **446,** no. 7133, 305–307.

Shelly, D. R., G. C. Beroza, and S. Ide (2007b). Complex evolution of transient slip derived from precise tremor locations in western Shikoku, Japan, *Geochem. Geophys. Geosyst.* **8,** Q10014, doi: 10.1029/2007GC001640.

Shelly, D. R., G. C. Beroza, S. Ide, and S. Nakamula (2006). Low-frequency earthquakes in Shikoku, Japan, and their relationship to episodic tremor and slip, *Nature* **442,** no. 7099, 188–191.

Wech, A. G., and K. C. Creager (2008). Automated detection and location of Cascadia tremor, *Geophys. Res. Lett.* **35,** L20302, doi: 10.1029/2008GL035458.

Ana C. Aguiar
Gregory C. Beroza
Department of Geophysics
Stanford University
Stanford, California 94305 U.S.A.
acaguiar@stanford.edu