32.1 Introduction

The normal modes spectra of the Earth’s free oscillations contain valuable information on the longest wavelength three dimensional structure of the Earth’s interior. Dahlen (1982) first showed that the optimum record length for measuring eigenfrequencies and decay rates of normal modes using a Hanning taper is 1.1 Q cycles. Since typical modal Q values lie in the range $10^2$ of to $10^3$, optimum window lengths are on the order of days to weeks. However, it is difficult even today to retrieve clean continuous time series spanning days or weeks following a large earthquake: in addition to aftershocks and other seismic events, data spikes due to transient disturbances at the station or temporary data storage failures are often unavoidable. Therefore, it is necessary to edit seismograms prior to the spectral analysis. Spikes are usually difficult to identify by existing algorithms because they have features which often are quite similar to those of real earthquake signals. The majority of conventional despiking algorithms require empirical tuning of associated parameters. The required parameter tuning often degrades the efficiency of automated algorithms, which in the end do not save much time and effort when compared with manual approaches. Hence, most despiking is done by direct manual editing, which can be a rather monotonous task, especially when one deals with a large dataset. Ideally, to be helpful, an automated despiking algorithm should not require users to manually intervene.

In this report, we describe a 3D phase space thresholding despiking algorithm which was first developed in the hydraulic engineering community and which we have adapted to our needs. The method was originally used to remove spikes from acoustic Doppler velocimeter data. It makes use of several ideas: (1) Differentiating a signal can enhance the high-frequency components (e.g. Roy et al., 1999); (2) The Universal Threshold (Donoho and Johnstone 1994b) introduced the Universal Threshold, which is given by

$$\lambda^U = \sigma \sqrt{2 \log n}$$

where $\sigma$ is the standard deviation of the noise sequence and $n$ is the number of data points.

(b) Differentiating and Calculating the Correlation Coefficients:

We calculate first and second derivatives of the original time series $\Delta u_i$ and $\Delta^2 u_i$ and the associated three sets of standard deviations $\sigma_u$, $\sigma_{\Delta u}$ and $\sigma_{\Delta^2 u}$. The correlation coefficients between $u - \Delta u$, $u - \Delta^2 u$ and $\Delta u - \Delta^2 u$ are calculated as follows:

$$\alpha_{u - \Delta u} = \tan^{-1} \left( \frac{\Sigma u_i \Delta u_i}{\Sigma u_i^2} \right) = 0;$$

$$\alpha_{u - \Delta^2 u} = \tan^{-1} \left( \frac{\Sigma u_i \Delta^2 u_i}{\Sigma u_i^2} \right);$$

$$\alpha_{\Delta u - \Delta^2 u} = \tan^{-1} \left( \frac{\Sigma \Delta u_i \Delta^2 u_i}{\Sigma (\Delta u_i)^2} \right) = 0;$$

(c) 3D Phase-Space Map:

Each set of three variables $\{u_i, \Delta u_i, \Delta^2 u_i\}$ determines a point $\{\rho, \theta, \phi\}$ in spherical coordinates, where $\rho^2 = u_i^2 + (\Sigma u_i)^2 + (\Sigma^2 u_i)^2$. For each pair of $\theta, \phi$, we can then calculate a threshold ellipsoid determined by:

$$\frac{1}{\rho^2} = \left( \frac{\sin \phi \cos \theta \cos \alpha + \cos \phi \sin \alpha}{a^2} \right)^2 + \left( \frac{\sin \phi \cos \theta \sin \alpha - \cos \phi \cos \alpha}{b^2} \right)^2 + \left( \frac{\sin \phi \sin \theta}{c^2} \right)^2$$

Where $a = \lambda_u^U = \sigma_u \sqrt{2 \log n}$, $b = \lambda_{\Delta u}^U = \sigma_{\Delta u} \sqrt{2 \log n}$ and $c = \lambda_{\Delta^2 u}^U = \sigma_{\Delta^2 u} \sqrt{2 \log n}$ and $\alpha$ is the rotation angle calculated from the correlation coefficient of $\alpha_{u - \Delta u}$.

The valid data points will then cluster inside of the threshold ellipsoids, while the data points that fall outside of the ellipsoids will be suspected as spikes.

32.2 Description of the Algorithm

The process of removing spikes comprises two steps: detection and replacement. In principle, these two parts are independent of each other, but the method described here is iterative and thus a proper spike replacing approach is also important for spike detection in the subsequent iterations.

(a) Universal Threshold:

As mentioned in the last section, Donoho and Johnstone (1994b) introduced the Universal Threshold, which is given by

$$\lambda^U = \sigma \sqrt{2 \log n}$$

where $\sigma$ is the standard deviation of the noise sequence and $n$ is the number of data points.

(b) Differentiating and Calculating the Correlation Coefficients:

We calculate first and second derivatives of the original time series $\Delta u_i$ and $\Delta^2 u_i$ and the associated three sets of standard deviations $\sigma_u$, $\sigma_{\Delta u}$ and $\sigma_{\Delta^2 u}$. The correlation coefficients between $u - \Delta u$, $u - \Delta^2 u$ and $\Delta u - \Delta^2 u$ are calculated as follows:

$$\alpha_{u - \Delta u} = \tan^{-1} \left( \frac{\Sigma u_i \Delta u_i}{\Sigma u_i^2} \right) = 0;$$

$$\alpha_{u - \Delta^2 u} = \tan^{-1} \left( \frac{\Sigma u_i \Delta^2 u_i}{\Sigma u_i^2} \right);$$

$$\alpha_{\Delta u - \Delta^2 u} = \tan^{-1} \left( \frac{\Sigma \Delta u_i \Delta^2 u_i}{\Sigma (\Delta u_i)^2} \right) = 0;$$

(c) 3D Phase-Space Map:

Each set of three variables $\{u_i, \Delta u_i, \Delta^2 u_i\}$ determines a point $\{\rho, \theta, \phi\}$ in spherical coordinates, where $\rho^2 = u_i^2 + (\Sigma u_i)^2 + (\Sigma^2 u_i)^2$. For each pair of $\theta, \phi$, we can then calculate a threshold ellipsoid determined by:

$$\frac{1}{\rho^2} = \left( \frac{\sin \phi \cos \theta \cos \alpha + \cos \phi \sin \alpha}{a^2} \right)^2 + \left( \frac{\sin \phi \cos \theta \sin \alpha - \cos \phi \cos \alpha}{b^2} \right)^2 + \left( \frac{\sin \phi \sin \theta}{c^2} \right)^2$$

Where $a = \lambda_u^U = \sigma_u \sqrt{2 \log n}$, $b = \lambda_{\Delta u}^U = \sigma_{\Delta u} \sqrt{2 \log n}$ and $c = \lambda_{\Delta^2 u}^U = \sigma_{\Delta^2 u} \sqrt{2 \log n}$ and $\alpha$ is the rotation angle calculated from the correlation coefficient of $\alpha_{u - \Delta u}$.

The valid data points will then cluster inside of the threshold ellipsoids, while the data points that fall outside of the ellipsoids will be suspected as spikes.
The 3D phase-space map has been used broadly in fractal geometry and chaotic dynamics studies. In phase-space, the transient high frequency component of the time series, which is much less random, can be separated from the chaotic regime that is featured as a compact cluster that is associated with chaotic oscillations. Because the time series needs to appear random for the Universal Threshold approach to be valid, the despiking procedure is not applied to the portion of data that are recorded within the first 21 hours after the origin time of the target event, during which strong coherent energy would not have the needed apparent randomness.

(d) Replacement:
Windows containing spikes can be considered as having gaps in the data stream. As one can already see, the shape of the threshold ellipsoid will vary when spikes are removed from a given segment of time series. Therefore, the entire despiking procedure needs to be repeated for several iterations until the number of detected spikes goes to zero. To ensure that the data cleaning procedure is complete and avoid introducing biases, it is important to develop a replacing strategy that can preserve the low frequency mode information contained in the time series.

The Discrete Fourier Transform (DFT) can be used to interpolate any data set that exhibits a periodic behavior. It consists of dividing the input signal into its major frequencies, determining the DFT coefficients (weights of each major frequency component), and then using the DFT coefficients and the associated frequencies to recompose the signal. This is the same process as is used for transmitting signals over telephone lines.

32.3 Application Examples
To explore the effectiveness of the automated despiking procedure while verifying that no significant bias is introduced during despiking, a set of spikes extracted from a noisy record is added to a raw time series free of spikes. The automated despiking algorithm is then applied on the artificially contaminated time series. We then compare the originally clean record with that obtained after applying the despiking procedure both in the time and frequency domain so as to check the preservation of the valid signal. Figure 2.69 shows the associated power spectra: The contaminated spectra (thick black line) and the post-despiking spectra (solid grey line) are plotted together. The striking effect of the added spikes is the severely elevated baseline level in the pre-despiking spectra. In Figure 2.69b, we compare the original un-contaminated power spectra (black dashed line) and the power spectra generated from the post-despiking time series (solid grey line). The fact that they are indistinguishable confirms that the despiking procedure is working well.

32.4 References