Geophysical Journal International

*Geophys. J. Int.* (2012) **189,** 602–610 doi: 10.1111/j.1365-246X.2012.05361.x

# Constructing a Hidden Markov Model based earthquake detector: application to induced seismicity

Moritz Beyreuther,[1] Conny Hammer,[2] Joachim Wassermann,[1] Matthias Ohrnberger[2] and Tobias Megies[1]

[1]*Department of Earth and Environmental Sciences (geophys. Observatory), Ludwig-Maximilians-Universität, Munich, Germany. E-mail: moritz.beyreuther@geophysik.uni-muenchen.de*
[2]*Institute for Earth and Environmental Sciences, Universität Potsdam, Potsdam, Germany*

Accepted 2011 December 29. Received 2011 December 27; in original form 2010 September 29

## SUMMARY

The triggering or detection of seismic events out of a continuous seismic data stream is one of the key issues of an automatic or semi-automatic seismic monitoring system. In the case of dense networks, either local or global, most of the implemented trigger algorithms are based on a large number of active stations. However, in the case of only few available stations or small events, for example, like in monitoring volcanoes or hydrothermal power plants, common triggers often show high false alarms. In such cases detection algorithms are of interest, which show reasonable performance when operating even on a single station. In this context, we apply Hidden Markov Models (HMM) which are algorithms borrowed from speech recognition. However, many pitfalls need to be avoided to apply speech recognition technology directly to earthquake detection. We show the fit of the model parameters in an innovative way. State clustering is introduced to refine the intrinsically assumed time dependency of the HMMs and we explain the effect coda has on the recognition results. The methodology is then used for the detection of anthropogenicly induced earthquakes for which we demonstrate for a period of 3.9 months of continuous data that the single station HMM earthquake detector can achieve similar detection rates as a common trigger in combination with coincidence sums over two stations. To show the general applicability of state clustering we apply the proposed method also to earthquake classification at Mt. Merapi volcano, Indonesia.

**Key words:** Time-series analysis; Neural networks, fuzzy logic; Seismic monitoring and test-ban treaty verification; Volcano seismology.

## 1 INTRODUCTION

For detecting earthquakes, counting associated triggers is a widely used method (e.g. short-time-average/long-time-average, STA/LTA; Withers *et al.* 1998). Especially when combining their output from multiple stations, their detection rates are high and the false alarm rates low. However, if there are only one or two stations available, the false alarm rate increases rapidly and, in the end, the seismologist usually has to interactively reanalyse the continuous data. In these cases, there is a strong demand for detection algorithms, which show good performance even when operating only on one or two stations.

In this study we focus on the detection of earthquakes which have similar characteristics, for example, similar frequency content and shape. Possible applications are reservoir monitoring, volcano monitoring or the detection of spatially clustered seismicity. To cope with the just described situation of sparse networks we design an earthquake detector with a probabilistic architecture which shows good performance when applied to a single station. The probabilistic architecture has been adopted from speech recognition and is known as Hidden Markov Model (HMM). In geoscience, HMMs are mainly known from the field of volcano seismology where they show good performance in classifying seismic data of volcanic origin (Ohrnberger 2001; Beyreuther *et al.* 2008; Ibáñez *et al.* 2009). To demonstrate their potential for seismology, Fig. 1 shows 9 hr of seismic data recorded at Mt. Merapi volcano including the colour coded earthquake types as classified by HMMs. Fig. 1 emphasizes the large potential of the HMMs: Even with a high number of different earthquake signals and signal amplitudes, HMMs are able to distinguish the target classes from each other (for more details on the Mt. Merapi volcano data set, Section 4.2).

Seismogram data from geothermal power-plants is dominated by anthropogenic noise as well as noise bursts and contains only a few induced seismic events. One reason for the high level of noise is the small distance of the power-plants to urban areas to which their operating company can sell the power and more important the heat. The data look often similar to that in Fig. 1, with the noise resembling the blue and green and the target waveform the red

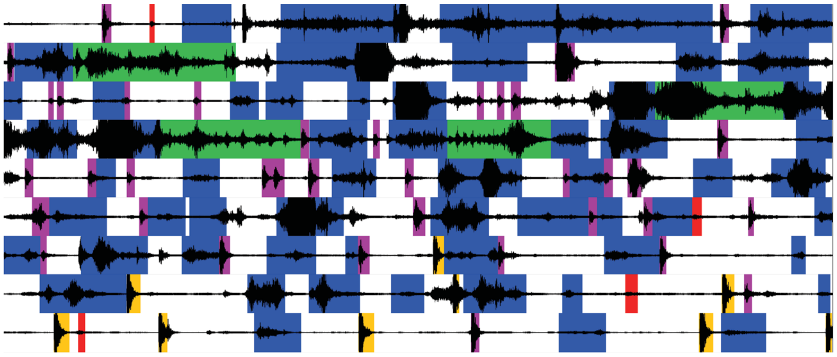

**Figure 1.** Classification results for 9 hr continuous recordings during an active period of the volcano Mt. Merapi. Colour codes correspond to Hybrids in magenta, block and ash flows in green, rockfall in blue, many phases events in red and Volcanic-Tectonic type B events in yellow and noise/without colouring. See Section 4.2 for a short description of the different classes.

or magenta earthquakes. This observation led to the idea to apply HMMs to detect induced seismicity.

However, when adopting HMMs from speech recognition to seismology several pitfalls need to be avoided. Especially incorporating the time dependent structure of the seismograms accurately in the model itself plays an important role as shown by Beyreuther & Wassermann (2011) who introduced Hidden Semi-Markov Models (HSMMs) to better model this time dependence. Unfortunately the detection and classification process using HSMM is extremely slow (1/2 hr CPU time for 1 hr data). As a way out of this dilemma, we approximate the time dependence of the HSMM with the faster HMM detection (40 s CPU Time for 1 hr data) by using state clustering. This approach allows also to introduce minimum duration length similar to the HSMM (Beyreuther & Wassermann 2011).

As a first step, we explain the meaning of the HMM parameters in the context of earthquake classification. In the remainder of the paper we concentrate on adapting the speech recognition models (HMMs) to seismology. In this step we introduce state clustering to model the time dependency of the HMM more adequately. The proposed HMMs are applied to continuous seismic monitoring data of a geothermal power-plant and the results are compared to those of a recursive STA/LTA trigger with coincidence sums over multiple stations. Hereinafter coincidence sums simply mean that a detection is only made, if the event was triggered for multiple stations in the same time span. To show the general applicability of both state clustering and the proposed adaptation steps, we show also the results of the application to the Mt. Merapi volcano 1998 data set.

## 2 FROM GAUSSIAN CLASSIFIERS TO HIDDEN MARKOV MODELS

The whole classification process consists of two steps. First, the HMM has to be set up through the analysis of preselected training data. This involves calculating a set of characteristic functions, transforming them to an orthonormal basis system and estimating Gaussian probability distributions and transition probabilities on the set of training observations. Then in the second step, these distributions and transition probabilities are used to estimate the best fitting event class in a sliding window on the continuous data that are to be analysed. To build a detector we simply recast the detection problem into a classification problem with two classes, noise for noise and induced for induced seismicity. In this section, we describe the meaning of the model parameters using a seismological example.

The seismic signal is not represented by the waveform itself but by a set of characteristic functions, which better discriminate the different classes. Many common triggering routines apply a similar approach, for example, the STA/LTA is working with simple triggering thresholds on one characteristic function computed from the waveform. Here we use for example, the characteristic functions ‘envelope’ and ‘instantaneous frequency’ (calculated in a running window) to discriminate the signal based on the coarse scale and on the small scale at the same time (for more details on characteristic functions and their selection process, see the studies on discrete HMM by Ohrnberger 2001; Beyreuther & Wassermann 2008). These characteristic functions are zero meaned and transformed to their principal axis. The mean and the transformation operator are pre-calculated from a principal component analysis of the training data. The resulting values are then inversely weighted by the square root of the associated singular value, yielding a multidimensional orthonormal time-series which is called features in this paper (Fig. 2). This whole procedure is also called pre-whitening, see Deller *et al.* (1993, p. 62) or Ohrnberger (2001, pp. 23–26). The transformation and subsequent normalization ensures the linear independence and unit variance of the resulting features.

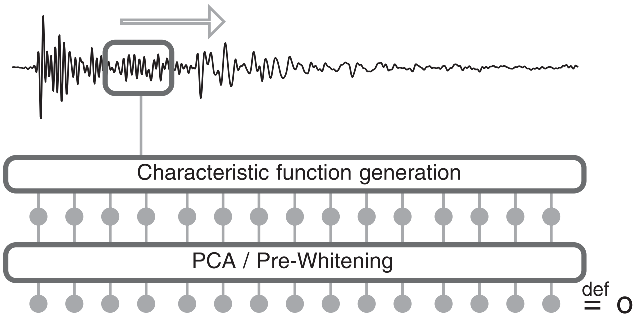
Characteristic function generation
PCA / Pre-Whitening
$\overset{\text{def}}{=} o$

**Figure 2.** Schema of the feature generation. The top seismogram is transformed in a sliding window to for example, 15 characteristic functions. The values of those characteristic functions (gray dots) are then transformed via the principal component analysis (PCA) and inversely weighted by the square root of the corresponding singular value. This methodology assures a linear independence and an unit variance of the resulting 15-dim feature vector $\boldsymbol{o}$ (represented by the bottom gray dots).

The features used in this study are shown in Fig. 3, where each feature is represented by one diagram. Each event in the training data set corresponds to one line in the various diagrams. The noise training samples are plotted in gray, the induced events in red. The vertical bars correspond to the estimated model parameters which we cover in more detail at the end of this section. Due to the transparency of the lines a crossing of multiple lines will result in a

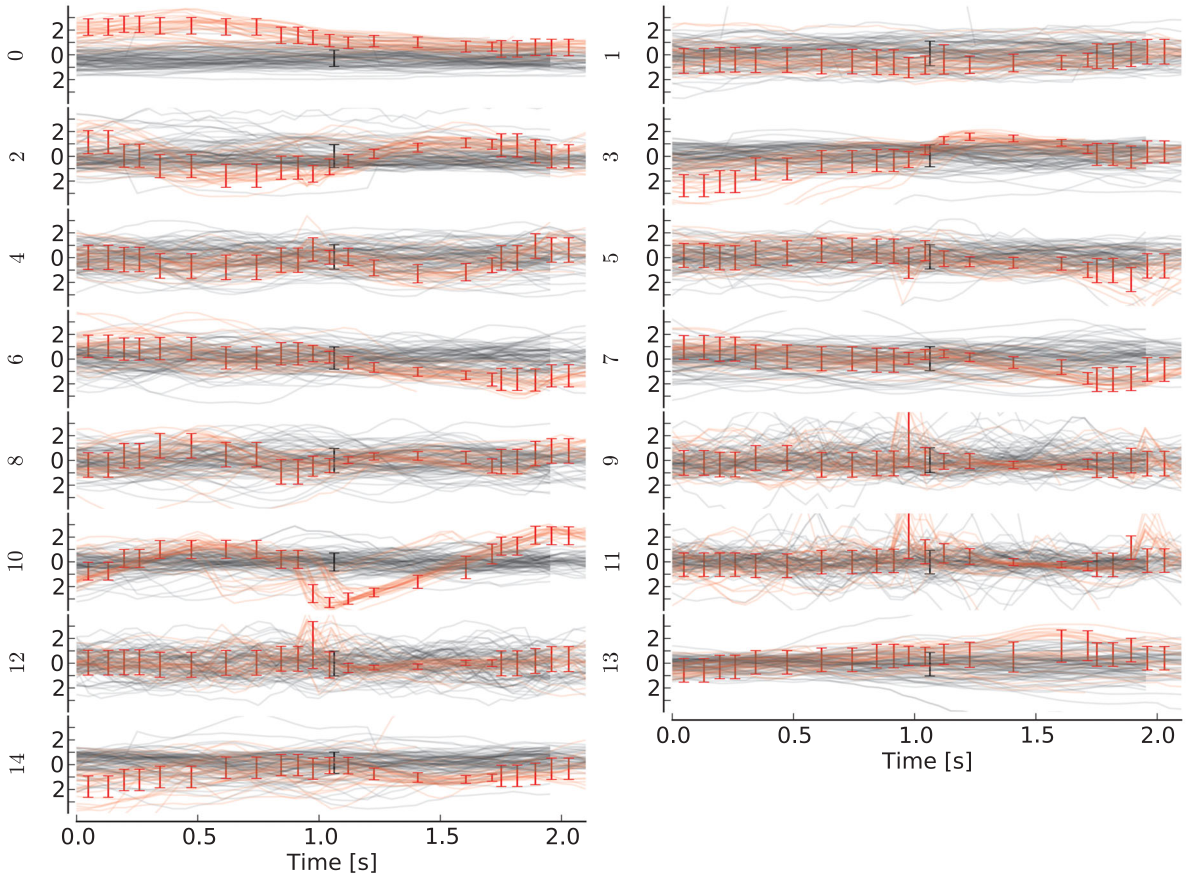

**Figure 3.** The different features used for HMM detection, where each feature is represented by one diagram. The number of each diagram corresponds to the 0th to 14th principal component of the 15-dim characteristic functions given in Table 1 normalized to unit variance. Each line corresponds to one event in the training data set, the induced earthquake events are shown in red, the noise events in gray. The 22 error-bars correspond to the trained 22 model states/Gaussians, some of which have variances tied (see the main text for details on the model representation).

more intense colour, which allows us to actually see the trend and the variability of the features. The extreme values are also visible, which would be omitted when solely plotting the mean and the variance or the median and the quantiles of the features, respectively.

We focus first on the feature labelled zero (top left-hand side in Fig. 3). By using all noise (gray) feature values (over time), we estimate one time independent Gaussian distribution $b$. The same procedure is repeated for the induced class in red. To classify one unknown feature value $o$, we choose that class, whose distribution $b$ has the highest probability for this unknown feature value $o$. For classifying multiple values over time $t$, we multiply the probabilities of every individual feature value in time $b(o_1)b(o_2)b(o_3)...b(o_t)$ and again the most probable class, that is, the one with the highest product, is chosen. The model parameters in this case are the mean and the variance of the noise and the induced distribution, respectively.

Obviously, the described technique works well for stationary signals, that is, the noise class in our case. Now we will focus on the feature labelled 10 in Fig. 3 and examine the induced earthquake class (red lines). The events are clearly time dependent, especially the feature values between 1 and 2 s change significantly in time. To estimate only one induced distribution $b$ for all values $o$ in time would lead to a blurred time dependent characteristic. Similar problems arise in spectral density estimation, where the common approach to avoid this effect is a time–frequency representation (e.g. short term Fourier transform or wavelet transform). A comparable idea is used here: Separate distributions $b_i$ are estimated for individual earthquake segments $i$, which are called states in the following. The different states are connected with state transition probabilities $a_{ij}$ (from state $i$ to state $j$), which are a measure for the duration time of state $i$. The combination of all those parameters, that is the means and variance of the Gaussians $b_i$ and state transition probabilities $a_{ij}$, is called HMM.

In practice the state transition probabilities $a_{ij}$ and the distributions $b_i$ are estimated jointly and automatically by an expectation maximization (EM) algorithm. Also, a single feature $o$ in practice is not enough to distinguish the classes. Therefore, the procedure is generalized and applied to multiple features, resulting in a $N$ dimensional distribution $b$ and feature vector $\boldsymbol{o}$.

Fig. 3 also shows the estimated probability distributions of $b_i$ after HMM training as error-bars. The centre of the error-bar represents the mean while the total height corresponds to twice the standard deviation. The horizontal distance between different error-bars represents the mean duration of a state as derived from the self transition probabilities $a_{ii}$. For a more formal and probabilistic description of HMMs, Rabiner (1989) and Young *et al.* (2002).

## 3 MODEL ADAPTATIONS FOR SEISMOLOGY

In speech recognition the training corpus/database is large (160 000 telephone utterances in Qing *et al.* 2000), which allows a robust estimate of even a large number of free HMM parameters. There exist similar attempts in volcano seismology to build up such databases. However, to this day often in earthquake classification only a low number of training earthquakes is available and statistically an over-fitting of the HMM parameters is likely. To nevertheless achieve robust results, the degrees of freedom of the HMM parameters ($a_{ij}$ and the means and variances of $b_i;\forall i, j$) have to be limited. It is crucial to refine the model parameters in such a way so that the HMM best represents the earthquake/event. A number of model adaptations is necessary, which are explained in the following subsections.

### 3.1 Left–right model topology

In their general definition HMMs allow transitions from any state to another. For instance in Fig. 3, a transition from the last error-bar back to the first error-bar is possible. However, a ‘beginning’ state will never follow the ‘coda’ state of the same earthquake and consequently the model topology is restricted to self transitions and transitions to the next state, (also called ‘left-right’ topology, Young *et al.* 2002, p. 107).

### 3.2 Grand variance models

The computation of a robust variance is the most difficult part in model estimation. Especially a low amount of training samples easily leads to an over-trained and thus useless model due to a small variance. The reason is that the training data set often does not represent all possible variations especially of noise in reality. Classifying 1 month of continuous data in a window with 4.5 s step size will result in approximately 500 000 classifications (3600 × 24 × 30/4.5) which will mostly be noise, and thus even a slightly over-trained noise model will lead to a high false alarm rate. Fig. 4(a) illustrates the problem by using an artificial feature. The light gray points correspond to induced earthquakes, the dark gray points to noise, respectively. For both classes a Gaussian is estimated with a mean feature value of 0.5 for the induced and 0.3 for the noise class. The most probable distribution for the unknown point at feature value 0.15 (denoted as a cross) is the induced distribution with the mean 0.5, even though the point is much closer to the mean/centre of the noise distribution at 0.3. This effect results from the normalization of the Gaussians to the area of one. Therefore, a smaller variance will yield a higher probability around the mean but a lower probability at the sides of the distribution (Fig. 4a). In case the noise distribution is over-fitted (the variance is underestimated, i.e. the modelled variance is smaller than the real variance) the point (cross) is actually misclassified as induced. One can avoid this behaviour by using grand variance models. Therefore the model variances are tied together. In the estimation tied variances behave like a single variance estimated from the training data of both classes, induced and noise. Fig. 4(b) shows the resulting grand variances. Note that the dark points corresponding to the noise class are certainly not described as well as by the also dark distribution in Fig. 4(a). Yet, the unknown point (cross) is classified correctly as the noise class to its closest Gaussian centre (Young *et al.* 2002, p. 152).

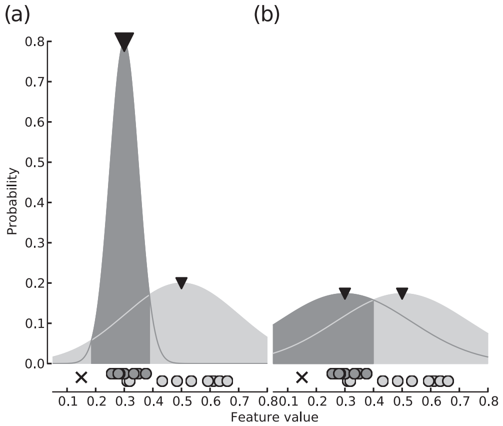

**Figure 4.** (a) Two Gaussians estimated from the dark and light gray points, the centres of the Gaussians are marked as black triangles. Even though the unknown point (cross) is closer to the centre of the dark gray distribution, the most probable distribution is the light gray one. This can be circumvented as shown on the right-hand side. (b) The two Gaussians are estimated with tied variances. The most probable distribution for the unknown point (cross) changes to the closest distribution centre, which is now the dark gray one.

However, some sections of the training data might be less prone to the over-fitting of noise. Here a better discrimination can be achieved by untying the variances of the induced class for those time sections. In the feature plot Fig. 3 the spread of the red lines of the individual feature values is quite high at the beginning and the end of the earthquake and low around 1–2 s (e.g. diagram 3 or 10 in Fig. 3). For the high spread the estimated high variance of the induced class would easily lead to concurring models for the noise class. Grand variances are used in this region. For the low spread of the red lines around 1–2 s the estimated variance of the induced class is low. In this case the induced class will be more likely than noise only in a small region around it’s mean, and thus the models do not concur. This observation is exploited by untying the variances of the induced class of the states corresponding to the 1-2 s time section.

### 3.3 State clustering

Although the state transition probabilities $a_{ij}$ and the distributions $b_i$ are estimated from the training data, the number of states to use is *a priori* set input parameter. A higher number of states allows a much better time discretization in the feature space, thus allowing a better description of the event. However, this also has the effect that the system is easily over-fitted (due to the high degree of freedom) and achieves poor results on data that differ slightly from the limited number of example data used for training.

Beyreuther & Wassermann (2011) showed that the correct time discretization has a large effect on the classification results when keeping the total number of states fixed. They propose HSMM in combination with Weighted Finite State Transducers which increased the classification rate significantly. However this method is extremely slow (1/2 h CPU time for 1 hr of data). Therefore, there is a strong demand for approximating a similar behaviour with a HMM approach (40 s CPU time for 1 h of data). To improve the time discretization without increasing the total number of states we propose to use state clustering, where states containing similar Gaussians (with respect to the Euclidean distance to their probability distributions’s mean) are tied together. In the training, tied parameters behave like a single one. Thus even if there is a high number of states, only the effective number of states after tying has an impact on the performance. The key parameter for clustering is the target number of clusters, in this case states, one wishes to have after clustering, which can be estimated in one of the following ways.

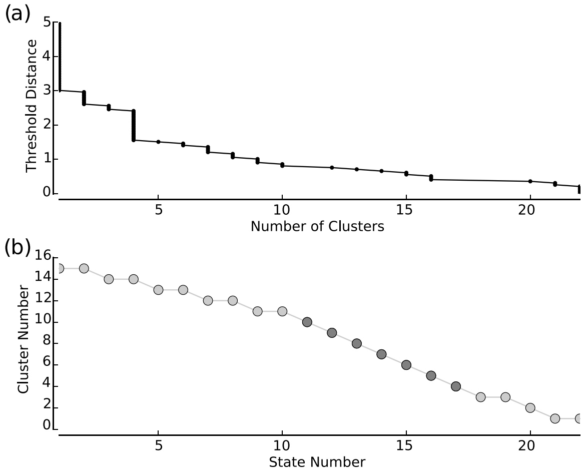

**Figure 5.** (a) Cluster threshold distance versus the number of states. (b) Cluster number versus total number of states: circles/states with the same cluster number are tied together. Light gray points correspond to states which have a grand variance, for the dark gray points the variances are not tied.

One possibility to select the number of clusters is based on a plot of the cluster threshold distance versus the number of states (Fig. 5a). The cluster threshold distance here means that only those states build a cluster (and are tied together), whose containing Gaussians (observation distributions) have an Euclidean distance between their centres that is less or equal the cluster threshold. Consequently a low cluster threshold will result in a high number of states. In our case we want to tie states containing similar Gaussians while keeping the prominent ones. From Fig. 5(a) we would select for example, four clusters.

However, using four clusters did not yield satisfying results. So the clusters were adjusted manually by counting the minimum number of states one would use to describe the earthquake in the feature space (Fig. 3) and using this as the target number of states. Based on Fig. 3 we decided to use 15 clusters.

An important effect of this clustering procedure is that it provides the possibility to introduce a minimum duration for a certain state. Fig. 5(b) shows for instance that the first state before clustering is represented by two states after clustering which are tied together (cluster number 15). In the new model, the state represented by cluster number 15 is entered twice and thus has a minimum length of two time samples. Of course also longer durations are allowed through the possible transition to the same earthquake state. However, the inherent problem becomes more obvious in this example: Consider a HMM which segments the earthquake in three states and a 20 time-samples long feature vector to classify. What we actually see in practice is that in the classification this state sequence is taken: $1 \rightarrow 2 \rightarrow 3 \rightarrow 3 \rightarrow 3 \rightarrow ... \rightarrow 3$. The whole classification collapsed to the third state. This can happen as the state transition probabilities $a_{ij}$ are of minor importance compared to the probabilities of the multidimensional distribution $b_i$ (i.e. the horizontal position of the error-bars in Fig. 3 varies more easily than the vertical position). A minimum duration length for each state avoids this collapse. It can be introduced by adding the same state multiple times. That is for the three state model a minimum duration of five time-samples per state can be introduced by designing a new 15 state model, consisting of five times state one, five times state two and five times state three, respectively. In the classification of the 20 time-sample long feature vector each of these states has to be entered at least once and thus only the last five time samples can collapse: $1_1 \rightarrow 1_2 \rightarrow ... \rightarrow 1_5 \rightarrow 2_1 \rightarrow ... \rightarrow 2_5 \rightarrow 3_1 \rightarrow ... \rightarrow 3_5 \rightarrow ... \rightarrow 3_5$.

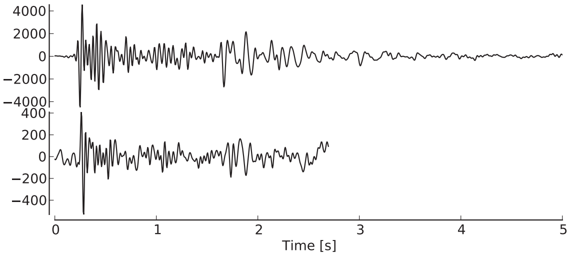

**Figure 6.** Two events whose length is determined such that all values are above signal-to-noise ratio two. Note that the upper event is twice as long as the lower one.

To avoid that the states of this new 15 state model are trained individually, effectively increasing the amount of free parameters, each five consecutive states are tied together (which allows to train them like a single state). Therefore, the clustering and the resulting tied states shown in Fig. 5(b) provide a data driven way to set the minimum length of each state, respectively.

### 3.4 Cutting the training data to similar length

In speech recognition the end of a certain word is quite clearly defined. In seismology the variability of the event-length for the same earthquake class is high and strongly depends on the coda. The coda contains backscattered seismic phases, which can be compared to echoes in the speech. These echoes are normally negligible in speech signals used for speech recognition. In seismology, however, these echoes/coda are strong and depend on the magnitude and the signal-to-noise ratio (SNR) of the earthquake. The larger the magnitude, the more waves are backscattered and the longer the coda and the events. The approach to select the length of the event based on the SNR is problematic as we demonstrate with the following example (Fig. 6): The median length of the training samples for the induced class is about 2.5 s, but there are high SNR samples in the training data set of up to 5 s, which additionally contain late coda. In Fig. 6 the upper event includes data samples of late coda which are drowned out and rendered unusable in the bottom event by the low SNR. The EM algorithm in the HMM training will include these samples in the estimation of the observation probability distributions $b_i$. Due to their long duration (2.5 s of late coda in the upper plot) these samples have a high impact on $b_i$. It will lead to a distorted $b_i$, which represents the coda of short events even worse. Thus the classification performance with respect to the short earthquakes will be much lower.

Consequently restricting the events to a maximum length allows a better performance especially in a low SNR environment. One could argue, that now the events containing late coda are represented worse. However, the following classifier is based on the confidence level ($P$(induced)/$P$(noise)) of a sliding [noise, induced, noise] and purely [noise] window. The late coda of high SNR events will then be classified as noise and is not taken into account.

### 3.5 Noise

As already described in Section 2, the time dependency is relevant only for the transient earthquake events. This time dependence is not needed in the case of the noise model, since noise most commonly does not have a causal or time dependent structure which could be described by different states. That is why we use a single state HMM (also called General Mixture Model) for the noise (e.g. Young *et al.* 2002, p. 34) and therefore the duration/event length of the noise training data (i.e. Section 3.4) has no primary effect on the classification result.

The exact parametrization of the noise and the induced class is given in the following section where the probabilistic earthquake detector is applied.

## 4 APPLICATION

### 4.1 Geothermal reservoir monitoring

The geothermal plant used as a case study here is located in the municipality of Unterhaching, which is a part of the district of Munich in southern Germany (Wolfgramm *et al.* 2007). To get better hypocentre estimates and to confirm the geothermal plant as the cause for the observed microseismicity, a local monitoring network was set up. In 2008 and 2009 two to three mobile stations were temporarily deployed in the epicentral region. Beginning in early 2010, five short-period seismometer stations were installed permanently at epicentral distances ranging from 2 to 9 km.

Here we apply the HMM based earthquake detector to recordings of one short period station (DHFO, Mark-L4) located 4 km south west of the geothermal plant. We think that this data set is particularly interesting for evaluating the performance for the following reasons:

(1) For monitoring geothermal power-plants that use existing hydrothermal reservoirs and do not rely on hydraulic stimulation of the reservoir usually just a few stations are used. Typically there are time intervals where some stations have recording gaps and/or are inactive due to technical difficulties, resulting in even less stations available for detection and classification. For the detection normally a STA/LTA in combination with coincidence sums is the method of choice. However, in cases where there are only two stations available, the coincidence sums will not be able to prevent noise triggers completely.

(2) For an earthquake detection and classification system, the false alarm rate is a crucial factor. Assuming a window step size of 4.5 s for the classification, in 1 month there are approximately half a million of classifications made (3600 × 24 × 30/4.5), which are mostly noise and contain just a few earthquakes. So a noise environment with all sorts of noise sources is needed for a realistic performance measure, especially noise events with a similar duration as earthquakes. This is given in particular near geothermal power-plants, which are usually located close to urban areas with the typical low SNR and all kinds of different noise sources (e.g. traffic, machinery, trees, pumps).

The two classes, induced earthquakes (induced) and noise (noise) served as examples in the theory section. The 15 characteristic functions used are given in Table 1 and their transformed counterpart (the 15 features) are shown Fig. 3. The trained HMM parameters are shown as error-bars in the same figure. In the following, we give a short description of the parameter constellation which we used.

**Table 2.** Geothermal data set. The table shows the number of events for training, their total amount of data-points/time samples and the median amount of data-points per event.

| Class | #Events | #DataPoints | Median(Data Points/event) |
|---|---|---|---|
| Induced | 21 | 1009 | 48.0 |
| Noise | 90 | 4180 | 45.0 |

Table 2 shows the number of training events and the corresponding amount of training samples. To better describe the time dependency, an initial induced earthquake model with 22 states is used (approximately half of the median of the data points/time samples per training event, see Table 2). These states are then clustered and tied to 15 independent states (Section 3.3), where each state contains a 15 dimensional Gaussian. As a consequence $15 \times 15$ means and variances need to be estimated from $1009 \times 15$ samples (15 due to the 15 dim Gaussians, see also Table 2), which results in just about 32 samples per mean/variance. For robust classification we rely on the grand variance models (Section 3.2), where the variance of the feature values is high (Fig. 3) and we only untie the variance of the middle part where the waveforms are really sharp and well defined (around 1–2 s in Fig. 3, Section 3.2). For the noise model, we use a single state HMM (Section 3.5). HMMs in general allow to use several Gaussian mixtures for each states. We tried a maximum of 10 Gaussian mixtures for the noise model but overall the results did not change dramatically. The total amount of trained parameters is for the noise class 15 means (from 15 features) plus one (tied global) variance. For the induced class the amount of trained parameters is 15 (from 15 independent states after clustering) times 15 means (from 15 features) plus 7 (from 7 untied states, corresponding to the 1–2 s period) times 15 variances (from 15 features) plus 22 transition probabilities (from 22 states) to the same state plus 21 transition probabilities to the next state.

For the continuous classification, a 9 s sliding window is classified with a step of 4.5 s. For each window a (noise, induced, noise) and a only (noise) sequence is classified, which allows to calculate the confidence level $P$(induced)/$P$(noise), with $P$ being the probability. In accordance with association rules (Hastie *et al.* 2009, p. 492ff) or the trigger thresholds we set the minimum log probability of a detection to $-16$ and a certain confidence level.

For the performance measure, continuous seismic data for the station DHFO was selected, starting in 2010 January. During this period two other stations were active as well which allowed us to verify the classification results by locating the events or, if that was impossible, by inspecting the waveforms at different stations. For the evaluation, the number of missed earthquakes is shown as a function of the false alarms (Fig. 7) for the HMM and a recursive STA/LTA trigger. To vary the false alarm rate, the confidence level is altered for the HMM from 0.01 on the right and 1 million on the left side of Figs 7(a) and (b). Similarly the 'trigger on' threshold for the recursive STA/LTA is increased from 2.7 on the right to 20 on the left side. The STA/LTA window length was set to 0.5 s and to 10 s for the STA and LTA window, respectively and the trigger off value is set to 1.5. The seismograms are first corrected to m/s and pre-filtered in three different frequency bands.

**Table 1.** Geothermal data set: a list of the used characteristic functions, separated by semicolons. *hob* corresponds to halve octave band, the frequency range is included in brackets. $Z$, $N$, $E$ correspond to the vertical, the north and the east component of the seismogram. The characteristic functions are transformed to their principal axis, the so-called features, which are plotted in Fig. 3. See Ohrnberger (2001); Beyreuther & Wassermann (2008) for the definition and a description of these characteristic functions.

*hob*1 (0.47–0.78 Hz); *hob*2 (0.70–1.2 Hz); *hob*3 (1.1–1.8 Hz); *hob*4 (1.6–2.6 Hz); *hob*5 (2.4–4.0 Hz); *hob*6 (3.6–5.9 Hz); *hob*7 (5.3–8.9 Hz); *hob*8 (8.0–13 Hz); *hob*9 (12–20 Hz); *hob*10 (18–30 Hz); normalized envelope $E$; instantaneous frequency $Z$; instantaneous frequency $N$; centroid time $E$; centroid time $N$.

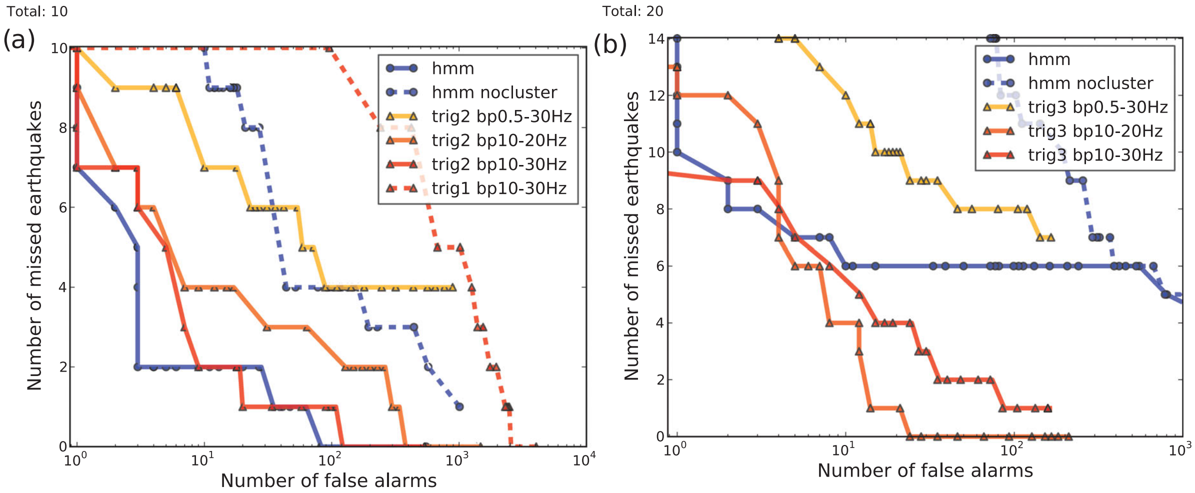

**Figure 7.** Geothermal data set: missed earthquakes as a function of false alarms for the HMM approach and a bandpassed trigger with coincidence sums. The labelling 'trig2 bp10-20Hz' corresponds to a trigger using two stations for coincidence sums where the data have been bandpassed between 10 and 20 Hz. (a) Corresponds to a 26 day period, where only two stations where active. (b) Corresponds to 3 month period where three stations were available.

In this 3.9 month period 2270400 classifications are made, whose results are split up in two periods. During the first period of 26 days, only two stations were active and the HMMs are compared to the trigger with coincidence sums over two stations in Fig. 7(a). On the top left hand side, the number of actual earthquakes in this time span is given as 10. For example, Fig. 7(a) shows that for two earthquakes missed (that is 8 of 10 are detected correctly) the HMM method has three false alarms (blue line), where the trigger bandpassed between 10 and 30 Hz (red line) has approximately 10 false alarms. Note that bandpassing the data before triggering in the frequency range 10–20 Hz (orange line) and 0.5–30 Hz (yellow line) decreases the performance of the trigger. The single station HMM method shows similar performance as the 10–30 Hz bandpassed trigger based on two stations and it clearly outperforms the trigger when operating on a single station (dashed red line). Using a 15 state HMMs directly without state clustering (blue dashed line; the remaining parameterization stayed the same) increases the false alarm by a factor of 10 to about 100 false alarms. This corresponds to an increase in the false alarm rate from about 0.002 per cent, to 0.02 per cent as in this first 26 days roughly 500 000 noise detections/classifications are made. Note that this small change in false alarm rate has an high impact on the detection performance.

Fig. 7(b) shows accordingly the results for the remaining 3 month period. During this period three and more stations were active. In these cases a single station method is more prone to miss those earthquakes which have very low SNR at this single station but higher SNR at the other stations. This behaviour can be observed in Fig. 7(b) where the three station trigger in red and orange clearly outperforms the HMM in blue for a low number of missed events. Nevertheless the HMM and the trigger operate with a similar performance at about 6–12 missed events. Note that again the HMM state clustering approach clearly outperforms the HMM without state clustering.

This application showed that the HMMs are helpful for detecting earthquakes when only one or two reference stations are available. The idea of using HMMs and the involved adaption steps were explained with this two class application. However, the real potential of HMMs is the classification of multiple classes, that is, one could introduce a quarry blast class and separate this from the induced earthquakes. However there are no quarry blasts in this data set. Therefore, and to show the general applicability of state clustering and the introduced model adaptation steps, we applied them to volcano monitoring, where multiple classes need to be distinguished. A short excerpt of this application is given next.

## 4.2 Volcano monitoring

To show the general applicability of state clustering we give more details on the classification system as shown in Fig. 1. Seismic surveillance at active volcanoes is one of the key areas for the application of automatic detection and classification. Here, we use data from a network which was installed at Mt. Merapi (Indonesia) in the years 1997–2006. During the analysed period of high activity in the year 1998 it consisted of four broadband seismic stations and 12 short-period stations which formed three arrays and one isolated station. Here, we only use the recordings of one broadband station (KLT, Streckeisen STS-2) located on the North West flank of the volcano. More details on the network can be found in Wassermann & Ohrnberger (2001). There are six different classes corresponding to Hybrids (HY), block and ash flows (PDF), rockfall (GU), many phases events (MP), Volcanic-Tectonic type B (VT) and noise. For a complete description of the signal characteristics please see the original paper of Minakami (1960) or Wassermann (2002). The size of the training data set is given in Table 3, the characteristic functions that were used are given in Table 4 (for more details on the features and their selection in case of Mt. Merapi, see Ohrnberger 2001, who applied Discrete HMMs to this data set). Similar to the main application, we construct a classifier. We use a class dependent window length: HY 51 s, PDF 383 s, GU 146 s, MP 52 s, VT 90 s. The window step is 18 s. The parameters for state clustering are given in Table 5. Four Gaussian mixtures were used for the noise

**Table 3.** Mt. Merapi data set: number of events used for training, their total amount of data-points/time samples and the median amount of data-points per event. The abbreviations correspond to Hybrids (HY), block and ash flows (PDF), rockfall (GU), many phases events (MP) and Volcanic-Tectonic type B (VT).

| Class Name | HY | PDF | GU | MP | VT | Noise |
|---|---|---|---|---|---|---|
| # Events | 32 | 19 | 73 | 30 | 30 | 65 |
| # DataPoints | 4364 | 15533 | 35 141 | 3478 | 4988 | 31 420 |
| Median (DataPoints/event) | 137 | 872 | 479 | 117 | 167 | 318 |

**Table 4.** Mt. Merapi data set. The used characteristic functions, separated by semicolons. *hob* corresponds to halve octave band, the frequency range is included in brackets. The characteristic functions are transformed to their principal axis, which are then called features and used as input for the HMM training. See Table 1 for a description of the characteristic functions.

*hob*1 (0.35–0.59 Hz); *hob*2 (0.52–0.88 Hz); *hob*3 (0.80–1.3 Hz); *hob*4 (1.2–2.0 Hz); *hob*5 (1.8–3.0 Hz); *hob*6 (2.7–4.4 Hz); *hob*7 (4.0–6.7 Hz); *hob*8 (6.0–10 Hz); normalized envelope $Z$; instantaneous bandwidth $Z$; centroid time $Z$.

**Table 5.** Mt. Merapi data set: cluster design. The total number of states per class is given in the first row. In the second row the numbers of clusters are given, which correspond to the number of independent states after clustering (see the main text for a definition of the abbreviations or Table 3).

| Class Name | HY | PDF | GU | MP | VT |
|---|---|---|---|---|---|
| # States | 52 | 325 | 190 | 44 | 64 |
| # Clusters | 13 | 8 | 10 | 17 | 12 |

distribution $b$ and the variances for PDF, GU and noise were tied together (grand variance). As an example the number of trained parameters for the HY class are: 13 (number of independent states after clustering) times 11 means and variances (from 11 features) plus 52 transition probabilities (due to 52 states) to the same state plus 51 transition probabilities to the next state.

The system was evaluated on 3.6 months of continuous seismic data. The outcome is presented in Fig. 8 which shows the results of the HMM classification versus the interactive event list of the local volcanologists. Note that especially the GU events match well the event rates classified by the local volcanologists. The MP events perform reasonable in general, and near the onset of the eruption especially well (green triangle in Fig. 8). However, there are still major differences in the VT and PDF rates. On the one hand the reason for this might be the low amount of training samples for the VT and especially the PDF class. More training data will allow to robustly estimate a higher number of independent states after clustering and thus provides an even more accurate time discretization. For the calculations we tested various characteristic functions described in Ohrnberger (2001) and selected the best subset. Nevertheless, the classification performance of the `VT` class could be improved by developing/adding another characteristic function which captures the main characteristics of the VT class. On the other hand the differences might be due to slight inconsistencies in the classification behaviour of the local volcanologists on the other hand. Note also, that the HMMs are operating on a single channel whereas the volcanologists compare the waveforms of several stations to set a certain class type. All training data are from time periods after 1998 July 1 (days containing training samples are marked as green rectangles in Fig. 8). Nevertheless the classification results from January to Mid-March still match the event rates of the volcanologists well. This is very promising as usually the characteristics of the events change slightly when the volcano is getting more active.

## 5 CONCLUSIONS

In this study we design a HMM based earthquake detector which operates on a single station. Despite of this restriction, the applied method has similar performance as a pre-filtered STA/LTA trigger with coincidence sums of two stations. We applied the detector to induced earthquakes from geothermal power-plants for a 3.9 month period. During most of this period three reference stations for the STA/LTA trigger were active and then the trigger outperforms the HMM technique. However, one of the reference stations was inactive for the first 26 d and immediately the false alarm rate of the STA/LTA trigger with only two stations for the coincidence sums rises dramatically. The STA/LTA performs not satisfactorily with

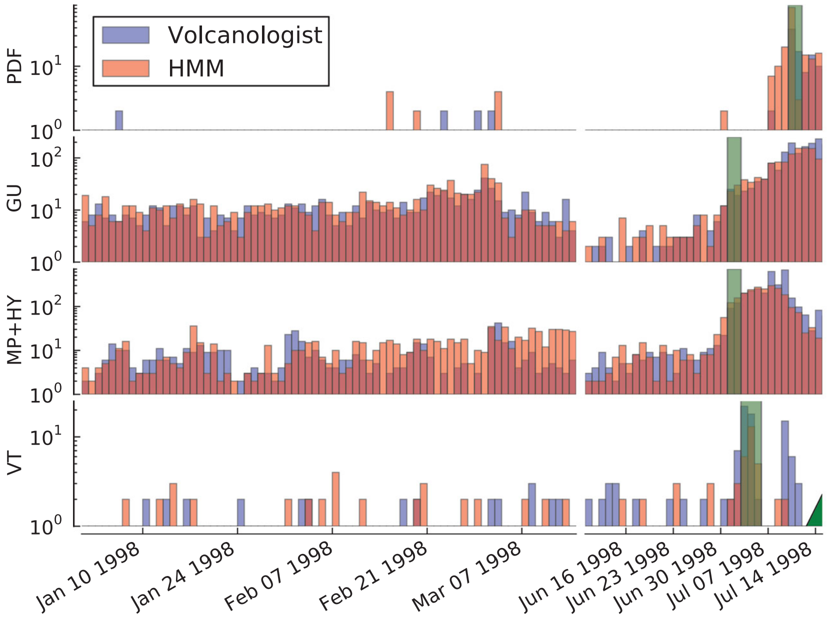

**Figure 8.** Mt. Merapi data set: continuous classification results. The light red bars correspond to the classification results of the HMM classification, the blue bars correspond to the classifications of the local volcanologists. The period starts on 1998-01-01, the station was inoperable between 1998-03-14 and 1998-06-10. The green triangle marks the onset of an eruption. The green rectangles mark days from which training data was taken. The bars are transparent, thus a blue bar overlaid with a light red bar results in a purple bar.

two stations available, and with only one station it is even impossible to extract the events because of nearly continuous triggering. For these cases the HMM based earthquake detector provides a promising and valuable alternative.

HMM in general are a technology borrowed from speech recognition. Even though earthquake signals look similar to speech signals, several pitfalls need to be avoided to apply the HMM in this field. HMMs have the advantage that the model parameters can be visualized to analyse the goodness of the fit. We provided an innovative way to show the model parameters (Fig. 3) together with the training data. This visual descriptive way of plotting the model parameters is an advantage compared to Neuronal Networks and Support Vector Machines whose parameters are usually associated with black boxes.

We further introduced state clustering to better model the time dependency of the HMMs. For the 3.9 month period of continuous described above, the false alarm rate increases at least by a factor of 10 when using HMM without state clustering. To show the general applicability of state clustering we applied the described system to seismic data of Mt. Merapi volcano. The resulting classification rates are promising and prove the general applicability of the proposed classification system. Especially the classifications of the Rockfalls (GU) is comparable with the interactive event list of the local volcanologists.

## ACKNOWLEDGMENTS

The authors wish to thank Heiner Igel for his kind support. The suggestions of two referees, Max Werner and one anonymous reviewer, substantially improved the manuscript. We also wish to thank the authors of the HMM Tool Kit (HTK) Young *et al.* (2002) whose powerful software package available at http://htk.eng.cam.ac.uk was used for the estimation of the model parameters and the Viterbi decoding. We released the recursive STA/LTA trigger used for the comparative study to the publicly available ObsPy toolbox (see Beyreuther *et al.* 2010). This work was partially funded by the German Ministry for Education and Research (BMBF), GEOTECHNOLOGIEN grant 03G0646H and by the Leibniz Institute for Applied Geophysics, contract number 201-4500053605.

## REFERENCES

Beyreuther, M. & Wassermann, J., 2008. Continuous earthquake detection and classification using Discrete Hidden Markov Models, *Geophys. J. Int.,* **175**(3), 1055–1066.

Beyreuther, M. & Wassermann, J., 2011. Hidden semi-Markov Model based earthquake classification system using Weighted Finite-State Transducer, *Nonlin. Processes Geophys.,* **18,** 81–89.

Beyreuther, M., Carniel, R. & Wassermann, J., 2008. Continuous Hidden Markov Models: application to automatic earthquake detection and classification at Las Canadas caldera, Tenerife, *J. Volc. Geotherm. Res.,* **176,** 513–518.

Beyreuther, M., Barsch, R., Krischer, L., Megies, T., Behr, Y. & Wassermann, J., 2010. ObsPy: a Python Toolbox for Seismology, *Seism. Res. Lett.,* **81**(3), 530–533.

Dellers, J.R., Proakis, J.G. & Hansen J.H.L., 1993. *Discrete-Time Processing in Speech Signals,* Prentice-Hall, Upper Saddle River, New Jersey.

Hastie, T., Tibshirani, R. & Friedman, J., 2009. *The Elements of Statistical Learning: Data Mining, Inference and Prediction,* 2nd edn, Springer, New York, NY.

Ibáñez, J.M., Benítez, C., Gutiérrez, L.A., Cortés, G., García-Yeguas, A. & Alguacil, G., 2009. The classification of seismo-volcanic signals using hidden markov models as applied to the stromboli and etna volcanoes, *J. Volc. Geotherm. Res.,* **187**(3–4), 218–226.

Minakami, T., 1960. Fundamental research for predicting volcanic eruptions (I) - Earthquakes and crustal deformations originating from volcanic activities, in *Bulletin of the Earthquake Research Institute,* Vol. 38, pp. 497–544, University of Tokyo, Tokyo.

Ohrnberger, M., 2001. Continous automatic classification of seismic signals of volcanic origin at Mt. Merapi, Java, Indonesia *PhD thesis*, Universität Postdam.

Qing, G., Yonghong, Y., Zhiwei, L., Baosheng, Y., Qingwei, Z. & Jian, L., 2000. Keyword spotting in auto-attendant system, in *Proceedings of the Sixth International Conference on Spoken Language Processing*. pp. 1050–1052, Beijing.

Rabiner, L.R., 1989. A tutorial on Hidden Markov Models and selected applications in speech recognition, *Proc. IEEE,* **77**(2), 257–286.

Wassermann, J. & Ohrberger, M., 2001. Automatic hypocenter determination of volcano induced seismic transients based on wave field coherence: an application to the 1998 eruption of Mt. Merapi, Indonesia, *J. Volc. Geotherm. Res.,* **110,** 57–77.

Wassermann, J., 2002. Chapter 13: Volcano Seismology, in *New Manual of Seismological Observatory Practise,* ed. Bohrmann, P., Deutsches GeoForschungsZentrum GFZ, Potsdam.

Wolfgramm, M. *et al.*, 2007. Unterhaching geothermal well doublet: structural and hydrodynamic reservoir characteristic, in *Proceedings of European Geothermal Congress,* Germany, Vol. 47, pp. 6–7.

Withers, M., Aster, R., Young, C., Beiriger, J., Harris, M., Moore, S. & Trujillo, J., 1998. A comparison of selected trigger algorithms for automated global seismic phase and event detection, *Bull. seism. Soc. Am.,* **88,** 95–106.

Young, S. *et al.*, 2002. *The HTK Book*. Technical report, HTK Version 3.2.1, Engineering Department, Cambridge University, Cambridge.