Geophysical Journal International

Geophys. J. Int. (2011) **185,** 253–264 doi: 10.1111/j.1365-246X.2011.04951.x

Clustering and classification of infrasonic events at Mount Etna using pattern recognition techniques

A. Cannata,[1] P. Montalto,[1,2] M. Aliotta,[1,3] C. Cassisi,[3] A. Pulvirenti,[3] E. Privitera[1] and D. Patanè[1]

[1] *Istituto Nazionale di Geofisica e Vulcanologia, Sezione di Catania, Piazza Roma 2, 95123 Catania, Italy. E-mail: montalto@ct.ingv.it*
[2] *Dipartimento di Ingegneria Elettrica, Elettronica e dei Sistemi, Università di Catania, Viale Andrea Doria 6, 95125 Catania, Italy*
[3] *Università degli studi di Catania, Dipartimento di Matematica e Informatica, Catania, Italy*

Accepted 2011 January 11. Received 2010 November 9; in original form 2010 July 28

SUMMARY

Active volcanoes generate sonic and infrasonic signals, whose investigation provides useful information for both monitoring purposes and the study of the dynamics of explosive phenomena. At Mt. Etna volcano (Italy), a pattern recognition system based on infrasonic waveform features has been developed. First, by a parametric power spectrum method, the features describing and characterizing the infrasound events were extracted: peak frequency and quality factor. Then, together with the peak-to-peak amplitude, these features constituted a 3-D 'feature space'; by Density-Based Spatial Clustering of Applications with Noise algorithm (DBSCAN) three clusters were recognized inside it. After the clustering process, by using a common location method (semblance method) and additional volcanological information concerning the intensity of the explosive activity, we were able to associate each cluster to a particular source vent and/or a kind of volcanic activity. Finally, for automatic event location, clusters were used to train a model based on Support Vector Machine, calculating optimal hyperplanes able to maximize the margins of separation among the clusters. After the training phase this system automatically allows recognizing the active vent with no location algorithm and by using only a single station.

Key words: Time series analysis, Volcano seismology, Volcano monitoring.

1 INTRODUCTION

Two of the fundamental tasks of volcano monitoring are to follow volcanic activity and promptly recognize any changes. To achieve such goals, reliable field measurements and advanced data analysis methods are required. Different geophysical techniques (i.e. seismology, ground deformation, remote sensing, magnetic and electromagnetic studies, gravimetry) are used to obtain precise measurements of the variations induced by an evolving magmatic system. In recent years, useful information to monitor the explosive activity of volcanoes, as well as to investigate its source processes, have been provided by studying infrasonic signals (e.g. Vergniolle & Brandeis 1994; Ripepe & Marchetti 2002; Cannata *et al.* 2009a,b; Marchetti *et al.* 2009). The location of the source of the infrasonic signals, generally coinciding with active vents, is of great importance for volcanic monitoring. Thus, different techniques, generally based on the comparison of the infrasonic signals using cross-correlation or semblance functions, have been developed (e.g. Ripepe & Marchetti 2002; Garces *et al.* 2003; Johnson 2005; Matoza *et al.* 2007; Jones *et al.* 2008; Montalto *et al.* 2010).

Over the last decades, Mt. Etna volcano (Italy) has been characterized by a remarkable increase in the frequency of short-lived, but violent eruptive episodes at the summit craters. Between 1900 and 1970, about 30 paroxysmal eruptive episodes occurred at the summit craters, while there have been more than 180 since then (Behncke & Neri 2003). The summit area of Mt. Etna is currently characterized by four active craters: Voragine, Bocca Nuova, Southeast Crater and Northeast Crater (hereafter referred to as VOR, BN, SEC and NEC, respectively; see Fig. 1). These craters are characterized by persistent activity that can be of different and sometimes coexistent types: degassing, lava filling or collapses, low rate lava emissions, phreatic, phreato-magmatic or strombolian explosions and lava fountains (e.g. Cannata *et al.* 2008). At Mt. Etna in 2006, a permanent infrasound network was deployed providing useful information to monitor the explosive activity (Cannata *et al.* 2009a,b; Di Grazia *et al.* 2009). Unfortunately, sometimes during the winter season owing to bad weather conditions, the lack of signals from some summit stations prevents applying the aforementioned location algorithms. Here, we propose a new system, based on pattern recognition techniques, able to identify at Mt. Etna the active summit crater from the infrasonic point of view using only the signal recorded by a single station.

2 INFRASOUND FEATURES AT Mt. ETNA

Some recent studies have shown that the infrasonic signal at Mt. Etna is generally composed of amplitude transients (named

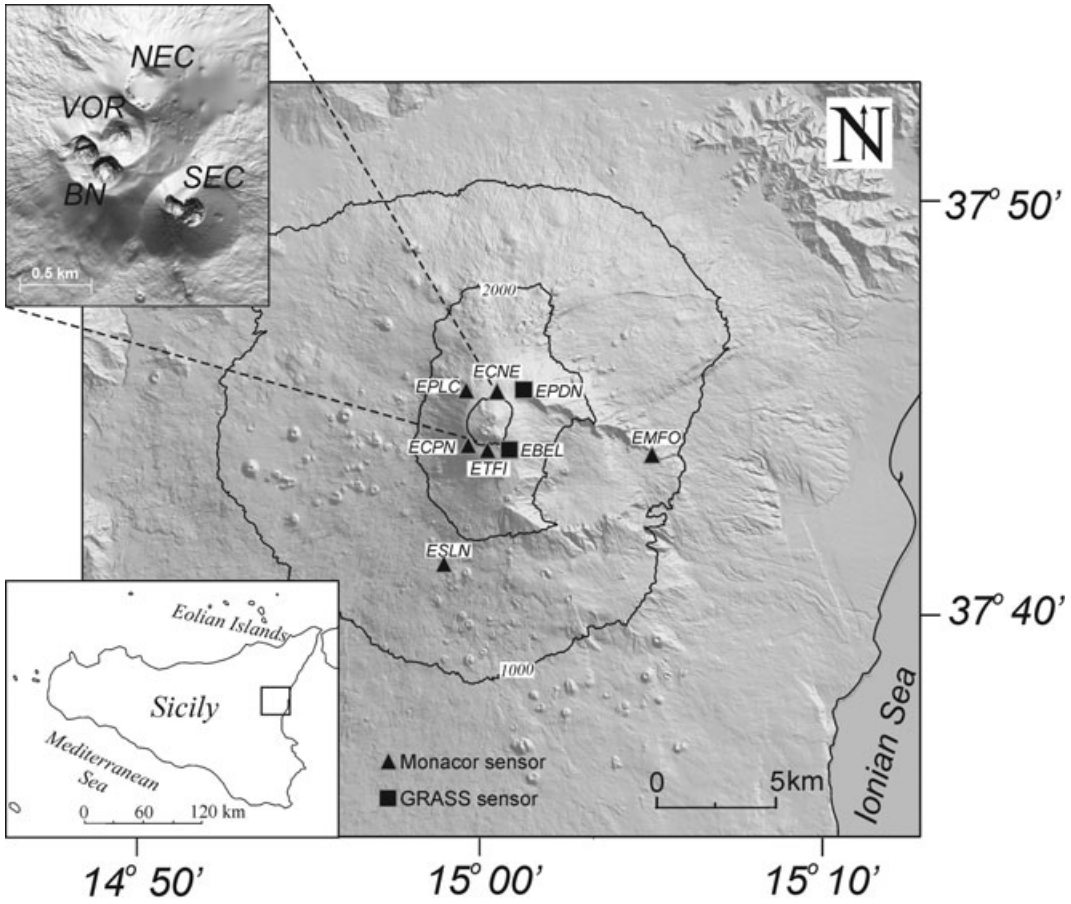

Figure 1. Digital elevation model of Mt. Etna with the location of the infrasonic sensors (triangles and squares), composing the permanent infrasound network. The upper right inset shows the distribution of the four summit craters (VOR, Voragine; BN, Bocca Nuova; SEC, Southeast Crater; NEC, Northeast Crater).

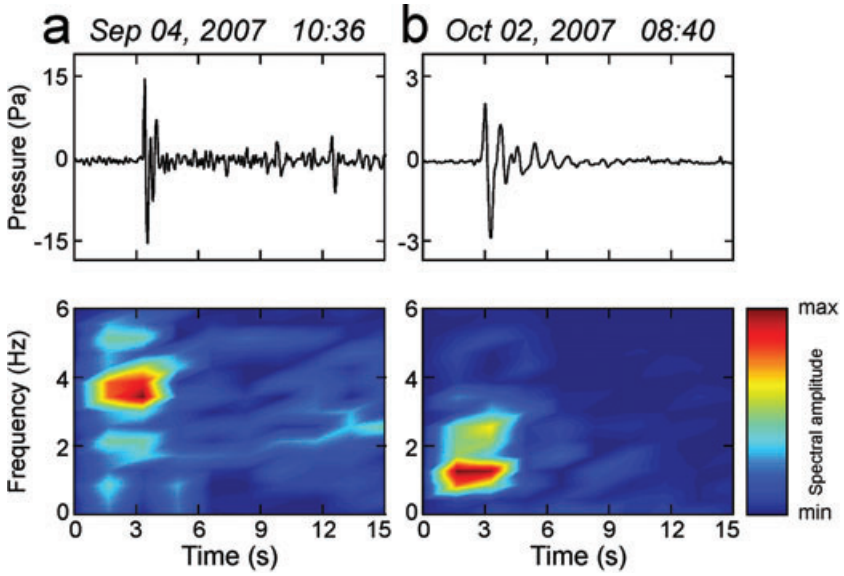

Figure 2. Infrasonic events recorded by EBEL station and corresponding Short Time Fourier Transform, obtained by using 2.56-s long windows overlapped by 1.28 s. The event in (a) is a typical 'SEC event', the one in (b) a typical 'NEC event'.

'infrasonic events'), characterized by short duration (from 1 to over 10 s), impulsive compression onsets and peaked spectra with most of energy in the frequency range 1–5 Hz (Fig. 2; Gresta *et al.* 2004; Cannata *et al.* 2009a,b). Similar features are also observed at several volcanoes, though characterized by different volcanic activity, such as Stromboli (Ripepe *et al.* 1996), Klyuchevskoj (Firstov & Kravchenko 1996), Sangay (Johnson & Lees 2000), Karymsky (Johnson & Lees 2000), Erebus (Rowe *et al.* 2000),

Arenal (Hagerty *et al.* 2000) and Tungurahua (Ruiz *et al.* 2006).

Since the deployment of the infrasound permanent network at Mt. Etna in 2006, two summit craters have been recognized as active from the infrasonic point of view: SEC and NEC (Cannata *et al.* 2009a,b). The former has been characterized by sporadic explosive activity with different intensity, from ash emission to lava fountaining, while the latter mainly by degassing. According to Cannata *et al*. (2009a,b), these craters generate infrasound signals with different spectral features and duration: 'SEC events', showing a duration of about 2 s, dominant frequency mainly higher than 2.5 Hz and higher peak-to-peak amplitude than the NEC events (Fig. 2a); 'NEC events', lasting up to 10 s and characterized by dominant frequency generally lower than 2.5 Hz (Fig. 2b).

3 DATA ACQUISITION AND INFRASOUND SIGNAL CHARACTERIZATION

In the following subsections (i) data acquisition and event detection, (ii) features extraction and (iii) the semblance algorithm are briefly described.

3.1 Data acquisition and event detection

Since 2006, the permanent infrasound network run by Istituto Nazionale di Geofisica e Vulcanologia, Section of Catania, has been composed of a number of stations ranging from one to eight depending on the considered period, located at distances ranging between 1.5 and 7 km from the centre of the summit area (Fig. 1). Today, some stations are equipped with Monacor condenser microphones MC-2005, with a sensitivity of 80 mV Pa^{-1} in the 1–20 Hz infrasonic band, while others with GRASS 40AN microphone with a flat response with sensitivity of 50 mV Pa^{-1} in the frequency range 0.3–20000 Hz. The infrasonic signals are transmitted in real-time by means of radio link to the data acquisition centre in Catania where they are acquired at a sampling rate of 100 Hz.

At Mt. Etna we use EBEL as reference station, because it generally shows a very good signal-to-noise ratio and, unlike the other summit stations, its maintenance is generally feasible even during the winter season. Once the infrasound signal is recorded, the signal portions of interest, that are the infrasonic events, have to be extracted. Then, the root mean square (rms) envelope of the infrasonic recordings is calculated by a moving window of fixed length. Successively, we calculate the percentile envelope on moving windows of rms envelope. For a given time-series, the pth percentile can be defined as the value such that at most $(100 \times p)$ per cent of the measurements are less than this value and $100(1 - p)$ per cent are greater. In light of this, the estimation of percentile enables us to efficiently detect amplitude transients and estimate background signal level. The percentage threshold should be chosen on the basis of both the amount of transients in the signal that have to be included or excluded in our calculations and the signal-to-noise ratio. The performance of this method was compared with the short time average/long time average (STA/LTA) technique (e.g. Withers 1997; Withers *et al.* 1998). The lengths of short and long windows, mainly depending on the frequency content of the investigated signal, were fixed respectively to 2.5 and 12.5 times the dominant period of the signal (equal to roughly 0.3 s), considered a reasonable compromise between sensitivity and noise reduction (Withers 1997), and the detection threshold to 1.7. As shown in Fig. 3, the trigger results obtained by the two methods were similar; nevertheless, the technique based on percentile was also able to detect transients very close to each other.

3.2 Infrasonic signal features extraction

Often the decomposition of a time-series into purely harmonic components (Fourier transform case) can be impractical. In fact, the actual oscillations observed in geophysics often decay (or grow) exponentially with time, due to some mechanisms of energy dissipation (or supply), as if the frequency were complex (Kumazawa *et al.* 1990). Therefore, the spectral structure will be reasonably represented in the complex frequency space (Kumazawa *et al.* 1990). Since infrasonic events can be represented as decaying complex exponential functions, to determine their complex frequency the Sompi method can be used (Kumazawa *et al.* 1990, & references therein). This is a high-resolution spectral analysis method based on an autoregressive (AR) filter. By this method, a given time series is resolved into a number of 'wave elements' that consist of decaying harmonic components, and additional noise (more details about Sompi method are reported in the Appendix). Each wave element is specified by two complex parameters z and α (Kumazawa *et al.* 1990)

$$z = \exp(\gamma + \mathrm{i}\omega) \tag{1}$$

$$\alpha = Ae^{\mathrm{i}\theta}, \tag{2}$$

where γ and ω are the real and imaginary parts of the complex angular frequency, A and θ correspond to the real amplitude and phase of the wave element referred to some origin point and finally i is $\sqrt{-1}$. Another two parameters, ordinary real frequency and 'gradient' or 'growth rate', referred as to f and g, respectively (Kumazawa *et al.* 1990), are given by

$$f = \omega/2\pi \tag{3}$$

$$g = \gamma/2\pi. \tag{4}$$

Finally, the 'dissipation factor' or 'quality factor' Q is defined as

$$Q = -f/2g. \tag{5}$$

Generally, to represent a set of complex frequencies, their locations are plotted on a 2-D plane with f and g axes. The wave elements scattering widely in the plot, as the AR order changes, are considered noise. It is also possible to identify some wave elements densely populated on the theoretical frequency lines that remain mainly stable as the AR order changes. They are considered dominant spectral components (Hori *et al.* 1989). An example of frequency-growth rate domain for an infrasound event recorded by EBEL station is reported in Fig. 4. Therefore, in summary, the spectral features of an infrasonic event can be described by the two parameters Q and f.

Further, in addition to frequency and quality factor, the third feature used to characterize the infrasound events is the peak-to-peak amplitude, depending on both distance source-station and energy of the infrasonic source.

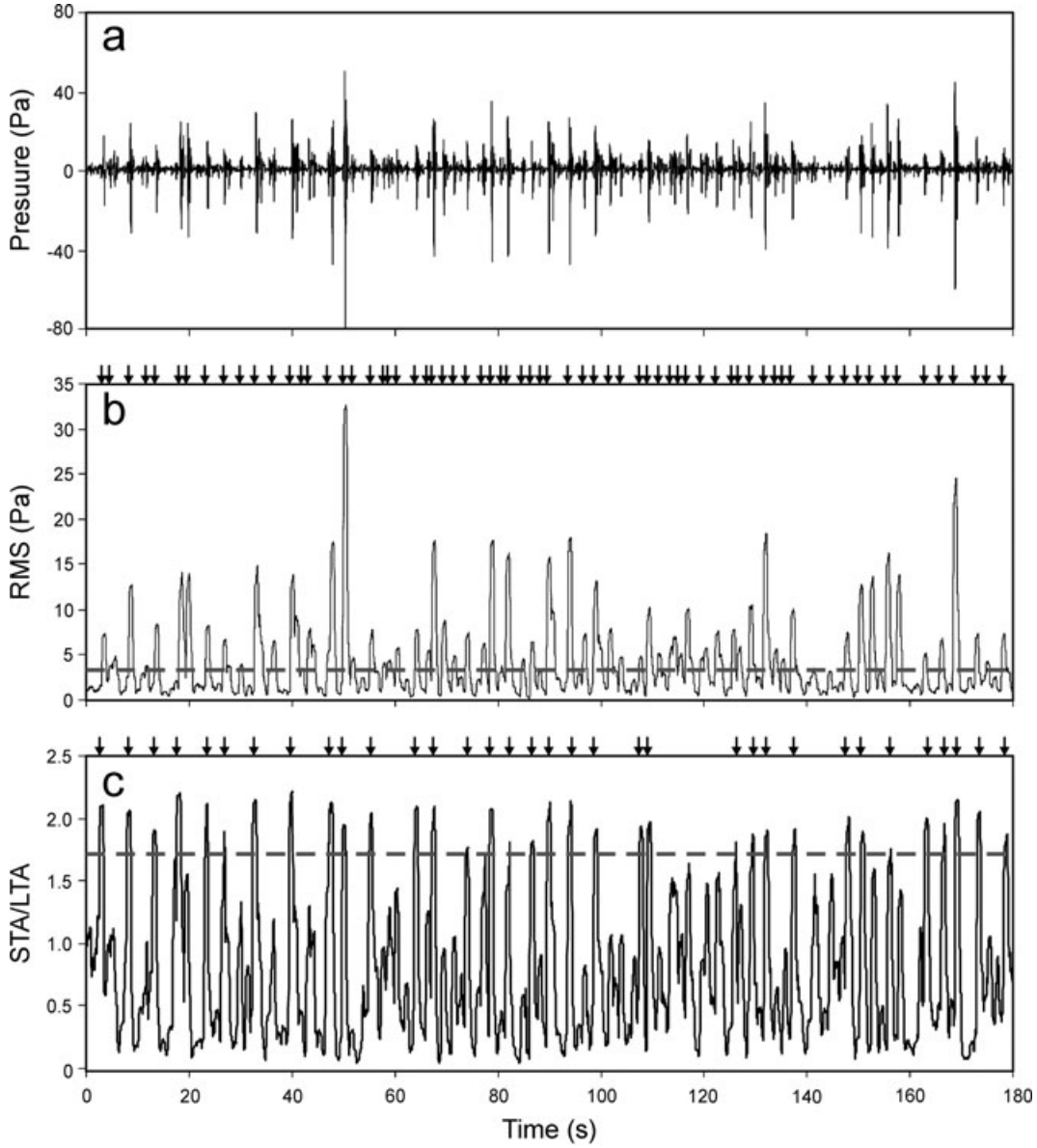

Figure 3. (a) Three-minute long infrasound signal recorded by EBEL station, (b) corresponding rms envelope (black line) calculated by using a moving window of 0.7 s and (c) STA/LTA values. The horizontal grey dashed line in (b) indicates the detection threshold calculated by a percentile value of 5 multiplied by 5. The horizontal grey dashed line in (c) indicates the detection threshold fixed at 1.7. The arrows at top of (b,c) indicate the onset time of the detected events.

3.3 Semblance algorithm

The location of the source of the infrasonic events, generally coinciding with active vents, is of great importance for volcanic monitoring. Therefore, different location techniques, generally based on grid searching procedures, were developed (e.g. Ripepe & Marchetti 2002; Jones *et al.* 2008; Johnson *et al.* 2010; Montalto *et al.* 2010). The semblance technique is based on the semblance function that is a measure of the similarity of multichannel data (Neidell & Taner 1971). For infrasonic events this method applies a 2-D grid searching procedure over a surface covering the summit area and coinciding with the topographic surface. The infrasonic source is assumed to be in each node of the grid, and for each node the theoretical traveltimes at the sensors are first calculated. Then, infrasonic signals at different stations are delayed and compared by the semblance function. Finally, the source is located in the node where the delayed signals show the largest semblance value. Therefore, the semblance function is assumed representative of the probability that a node has to be the source location (further details about the method are reported in Montalto *et al.* 2010). In Fig. 5 two examples of infrasound location are reported for a SEC event and a NEC event.

4 PATTERN RECOGNITION TECHNIQUES

Automatic extraction, recognition, description and classification of patterns extracted from images and signals are important tasks in several scientific disciplines. For instance, several studies based on pattern recognition (hereafter referred to as PR) techniques have been performed on seismo-volcanic signal analysis (e.g. Ohrnberger

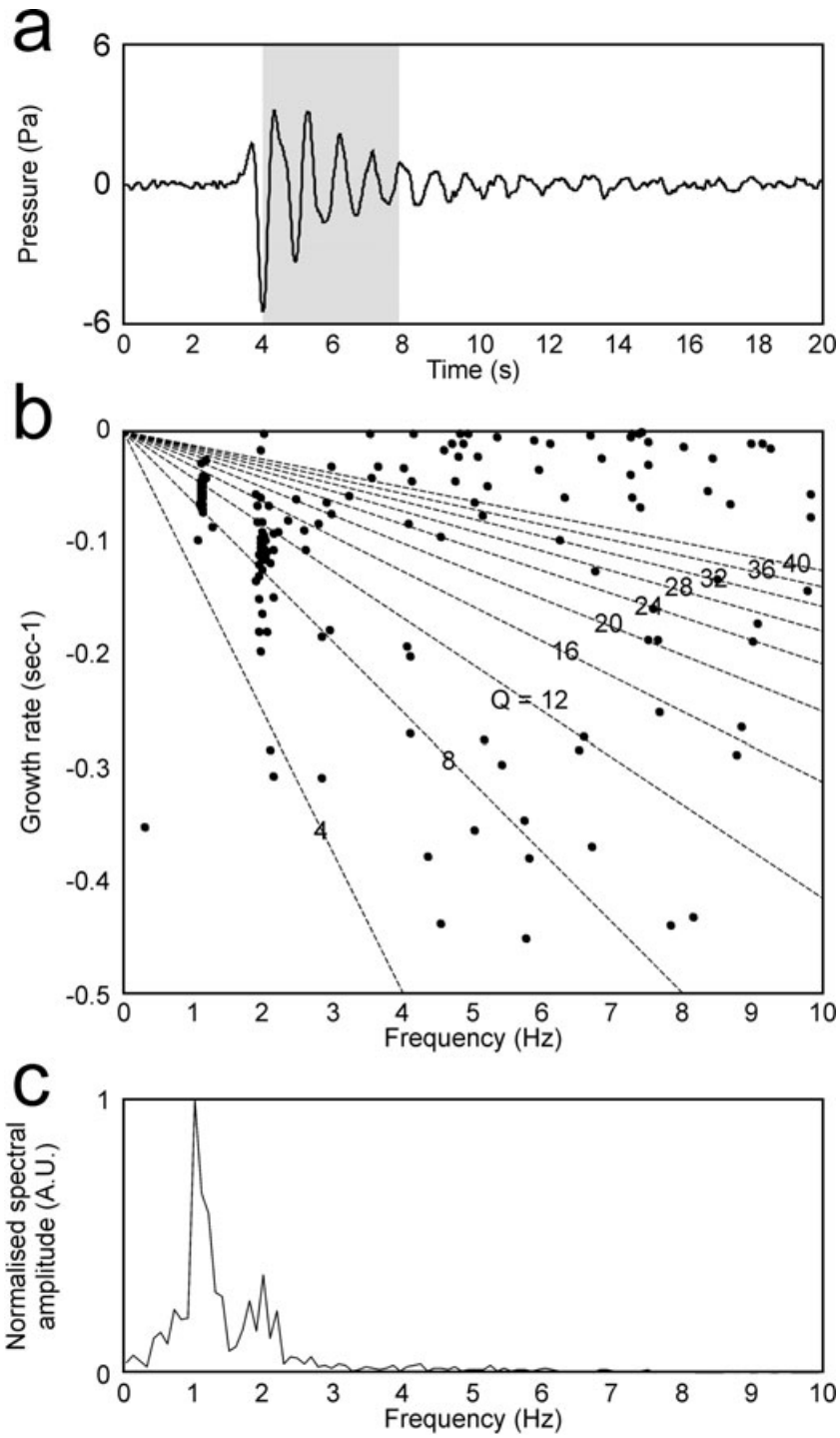

Figure 4. (a) Infrasound event recorded by EBEL station, and corresponding (b) frequency-growth rate plot (AR order 2–60) and (c) amplitude spectrum. The grey area in (a) represents the window used to calculate the frequency-growth rate plot in (b). The dashed lines in (b) represent lines along which the quality factor (Q) is constant. Clusters of points in (b) indicate dominant spectral components of the signal; scattered points represent noise.

2001; Kohler *et al.* 2009). The main aspect of PR is the definition of a set of peculiar features or descriptive elements of the analysed objects. Given a pattern, the recognition process consists of one of the following tasks: (i) supervised classification in which objects are classified on the basis of inference rules acting on a set of knowledge patterns (Joswig 1990); (ii) clustering that is the process of grouping sets of objects into classes called clusters with no a priori knowledge.

In the first case, the classifier design can be implemented by several techniques, implying the definition of a metric based on template matching or the minimum distance between pattern and class prototype. Other classification techniques are based on geometric approaches. These kinds of classifiers are based on a training procedure that minimizes an error (such as the mean square error, MSE) computed comparing classification output and target value. A powerful method in classifier design is the Support Vector Machine

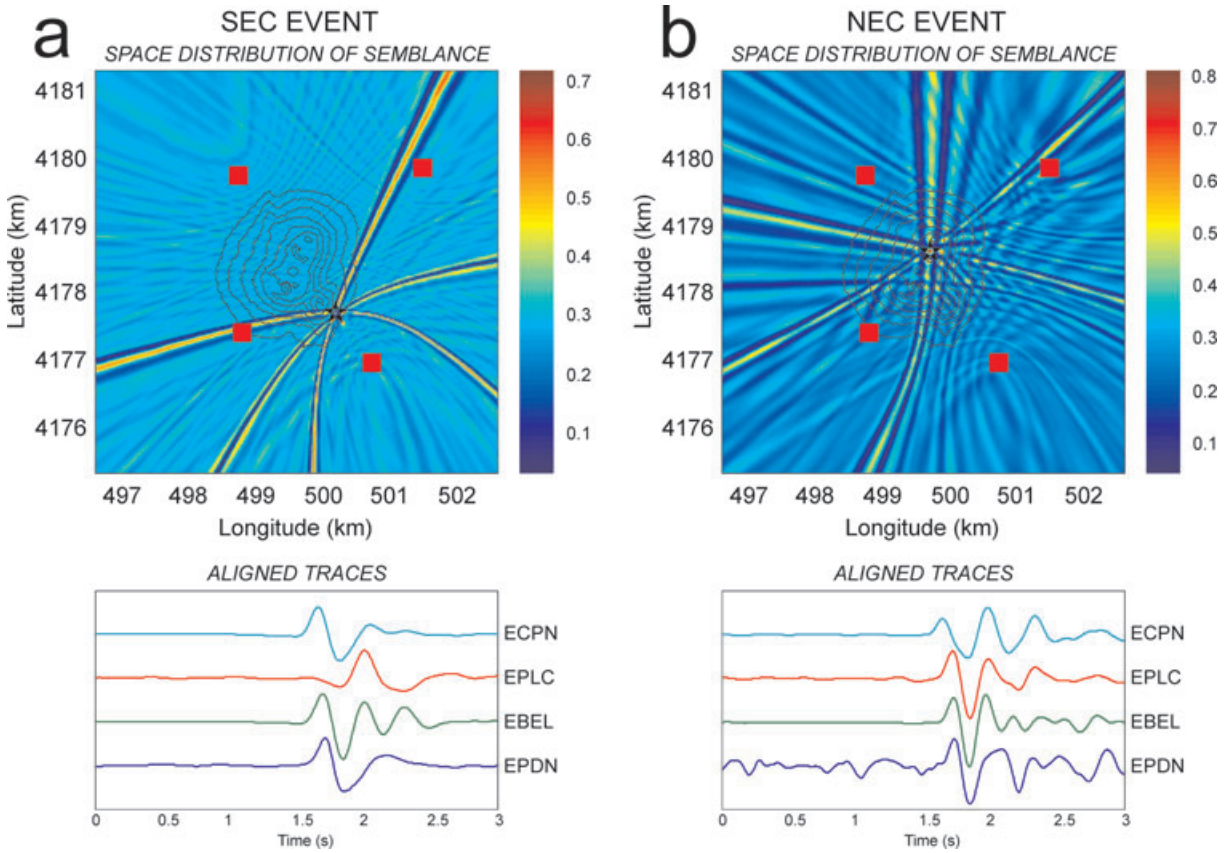

Figure 5. Examples of space distribution of semblance values, calculated by locating two infrasonic events at Mt. Etna, and corresponding infrasonic signals at four different stations shifted by the time delay that allows obtaining the maximum semblance. The red squares and stars in the top plot indicate four station sites and the nodes with the maximum semblance value, respectively. The black lines in the top plot are the altitude contour lines from 3 to 3.3 km a.s.l.

(SVM) introduced by Vapnik (1998). This algorithm is different from other hyperplane-based classifiers such as single layer perceptron. The problem of estimating hyperplane that separates two classes is not unique. The SVM algorithm is able to find the optimal hyperplane that separates the classes.

Another important task in PR is the clustering problem that, as aforementioned, is the process of grouping data without any a priori information. Objects belonging to the same cluster will be more similar than objects belonging to different clusters with respect to some given similarity measures. Many clustering methods exist in literature (Berkhin 2002). These can be broadly divided into hierarchical and partitioning. Hierarchical algorithms gradually (dis)assemble objects into clusters. On the other hand, partitioning algorithms learn clusters directly, trying to discover clusters either by iteratively relocating points between subsets or by identifying areas heavily populated with data. This second type of partitioning algorithms attempts to discover dense connected components of data. Examples of algorithms belonging to such a category are: DBSCAN, OPTICS and DENCLUE (Berkhin 2002). The system proposed in this work uses both clustering and classification algorithms to develop an automatic procedure able to discover and classify clusters in a given feature space. The algorithm chosen for pattern clustering and classification are DBSCAN and SVM, respectively.

4.1 Clustering algorithm based on DBSCAN

Density-Based Spatial Clustering of Applications with Noise (DBSCAN; Ester *et al.* 1996) is a density-based clustering algorithm able to discover clusters of arbitrary shape in spatial databases with noise. Clusters are defined as maximal sets of density-connected points. Usually DBSCAN runs on data sets drawn from multidimensional or metric spaces and uses a distance function to compare objects. Given a data set D of objects, DBSCAN makes use of the following structures and definitions (Ester *et al.* 1996): (i) ε-neighbourhood, (ii) core point, (iii) directly density-reachable, (iv) density-reachable and (v) density-connected. The ε-neighbourhood of a point p, denoted by $N_\varepsilon(p)$, is a subset of points q in D, such that a distance measure dist(p,q) (such as the Euclidean distance) is lower than ε. The point p is called core point or core object if its ε-neighbourhood has cardinality above a minimum threshold called *MinPts*. Each point q which lies in the ε-neighbourhood of a point p is called directly density-reachable from p (Fig. 6a). A point q is density-reachable from a point p with respect to ε and *MinPts* if there is a chain of points $q_1,\ldots,q_n$ such that $q_1 = p$, $q_n = q$ and q_{i+1} is directly density-reachable from q_i for each i (Fig. 6b). A point q is density-connected to a point p with respect to ε and *MinPts* if there is a point o such that both p and q are density reachable from o with respect to ε and *MinPts* (Fig. 6c).

Given D, ε and *MinPts* as input parameters, DBSCAN clusters D by checking the ε-neighbourhood of each object in D. If the ε-neighbourhood of an object p contains more than *MinPts*, a new cluster with p as core object is created. DBSCAN iteratively collects directly density-reachable objects from these core objects. The process terminates when no new objects can be added to any cluster. In such a case, the algorithm will return the set of clusters and a special cluster containing outliers.

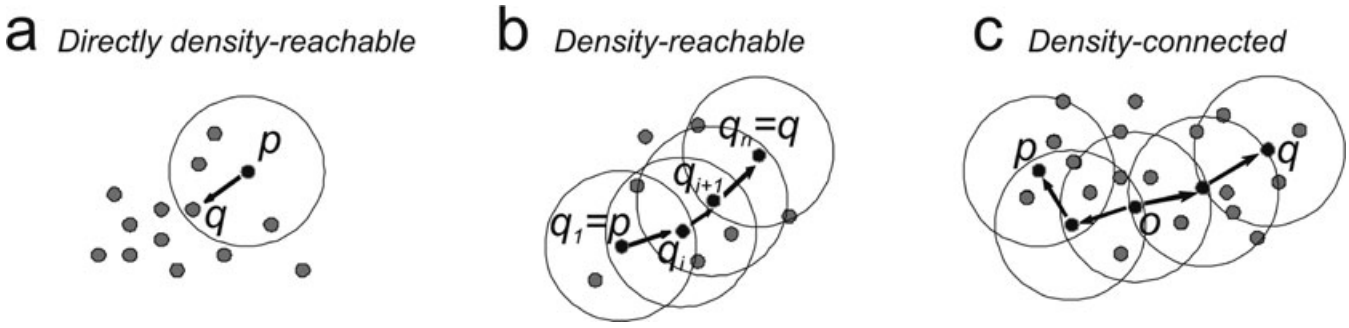

Figure 6. Examples of (a) directly density-reachable, (b) density-reachable and (c) density-connected in density-based clustering. Suppose *MinPts* = 3. Grey and black dots indicate the points to group into clusters, black circles delineate the area of radius ε around black dots, the arrow denotes the relation of direct density-reachability. In (a) dot p is the so-called core point, while q is directly density-reachable from p. In (b) dot q is density-reachable from p. In (c) dot q is density-connected to p and o is a point such that both p and q are density reachable from o (for details see Section 4.1).

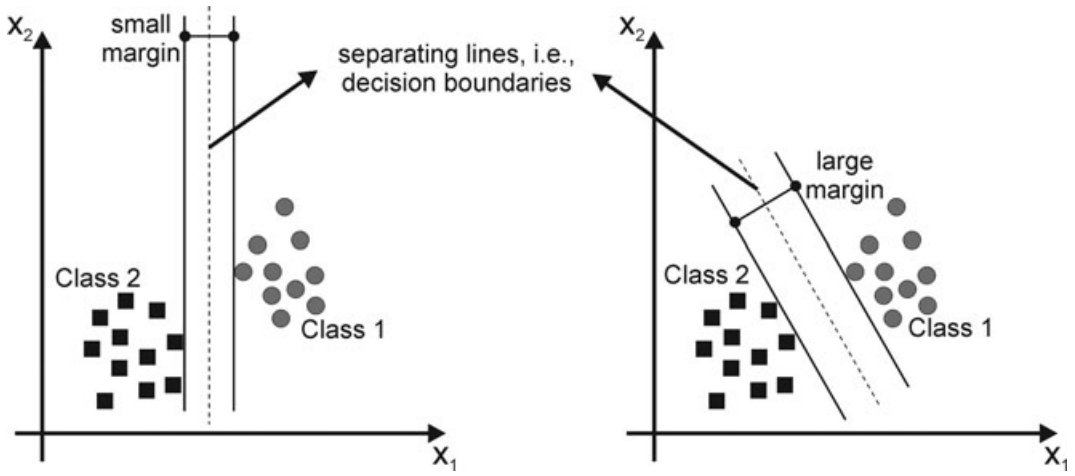

Figure 7. Two-feature planes each of which with two classes of data (black squares and grey circles) and a separating line (dashed lines): the left one shows a small margin between clusters, the right one a larger margin (redrawn from Kecman 2001).

To estimate the best clustering structure we used an internal cluster validation measure called Davies–Bouldin (DB) index (Davies & Bouldin 1979). This index is a function of the number of clusters, the intercluster and within-cluster distances. Formally it is defined as follows

$$\mathrm{DB} = \frac{1}{n}\sum_{i=1}^{n} \max_{i\neq j}\left\{\frac{S_n(Q_i)+S_n(Q_j)}{S(Q_i,\,Q_j)}\right\}, \tag{6}$$

where n is the number of clusters, S_n is the average distance of all cluster objects to their cluster centre, $S(Q_i,\, Q_j)$ is the distance between clusters centres. Small values of DB correspond to compact clusters whose centres are far away from each other. Thus, the number of clusters that minimizes DB is taken as the optimal number of clusters. Although such a clustering phase may prove expensive (i.e. run the algorithm several times with different parameters), this is performed only once. In the second phase, the cluster containing outliers (i.e. the noise) is removed from the data set.

4.2 Features classification using SVM

SVMs are a popular machine learning method for solving problems in classification and regression, able to guarantee high classification quality (Burges 1998). In recent years, novel applications of SVM have been performed in several research areas such as biology (e.g. Noble 2004; Cheng *et al.* 2006) and volcano seismology (e.g. Masotti *et al.* 2008; Langer *et al.* 2009). The SVM algorithm can be summarized as follows. It first uses a non-linear mapping to transform the original data set into a higher dimension space. Next, it identifies a hyperplane able to maximize the margin of separation among the classes of the training set. Such a hyperplane is called maximum marginal hyperplane (MMH). The margin in SVMs denotes the distance from the boundary to the closest data in the feature space (Fig. 7). With appropriate mapping, data from two classes can always be separated by a hyperplane. The problem of computing the MMH can be formulated in terms of quadratic programming in the following way (Hwanjo *et al.* 2003).

$$W(\alpha) = -\sum_{i=1}^{l}\alpha_i + \frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l} y_i y_j \alpha_i \alpha_j k(x_i, x_j) \tag{7}$$

subject to

$$\begin{aligned} &\sum_{i=1}^{l} y_i\alpha_i = 0 \\ &\forall i : 0 \le \alpha_i \le C. \end{aligned} \tag{8}$$

The number of training data is denoted by l, α is a vector of l variables, where each component α_i corresponds to a training data (x_i, y_i). C is the soft margin parameter controlling the influence of the outliers (or noise) in training data.

The kernel for linear boundary function is $x_i y_i$, a scalar product of two data points. The non-linear transformation of the feature space is performed by replacing $k(x_i, y_i)$ with an advanced kernel φ, such as polynomial kernel $(x^T x_i + 1)^p$ or a radial basis function kernel

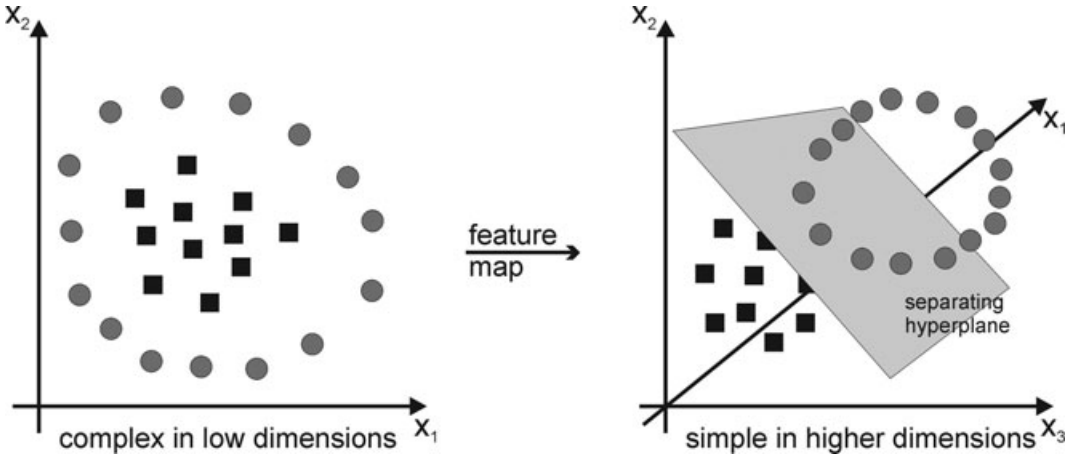

Figure 8. Two classes of data in the original 2-D space (left) and in a higher-dimensional feature space (right).

$\exp(-\frac{1}{2\sigma^2}\|x - x_i\|^2)$. The use of an advanced kernel is an attractive computational shortcut, which avoids the expensive creation of a complicated feature space. An advanced kernel is a function that operates on the input data but has the effect of computing the scalar product of their images in a usually much higher-dimensional feature space (or even an infinite-dimensional space), which allows one to work implicitly with hyperplanes in such highly complex spaces (Fig. 8). The extension of SVM to multiclass problems can be performed using two different methods called one-against-one and one-against-all. The former constructs $k(k-1)/2$ classifier where each one is trained on data from two classes. The latter constructs k SVM classifier. In this last case, the ith SVM is trained using all training patterns belonging to ith class with positive labels and the other with negative labels. A point is assigned to the class for which the distance from margin is maximal. Finally, the output of one-against-all method is the class that corresponds to SVM with highest output value (Weston & Watkins 1999; Hsu & Lin 2002).

4.3 Learning phase

In the proposed system, the learning phase merges together results of clustering and classification analysis (Fig. 9). The techniques described in Section 4.1 and 4.2 are applied on infrasound event features together with geophysical information used to 'label' the recognized clusters. About 665 events, recorded during 2007 September–November at EBEL station, were detected and filtered in frequency range 0.5–5 Hz. The feature extraction from the detected events was performed by Sompi method (Section 3.2) using 2-s long windows of infrasonic signal recorded at EBEL station and AR order equal to two. The sharply monochromatic nature of the investigated signals justifies the choice of this low order (Lesage 2008). Frequency and quality factor of the events, together with peak-to-peak amplitude, constituted the feature space and are plotted in Fig. 10. Then, to discover clusters in this space, 'data clustering' techniques based on DBSCAN algorithm (Section 4.1) were applied. Using such an algorithm we found three main clusters (called cluster 1, 2 and 3) and other outlier points that can be considered as noise (Fig. 11). Points belonging to each cluster are related to infrasonic events that were located using Semblance location method (Section 3.3). In accordance with Cannata *et al.* (2009b), during 2007 September–November, two infrasonic sources were found, NEC and SEC. In particular, a cluster was composed of events generated by NEC (cluster 1) and the other two by SEC.

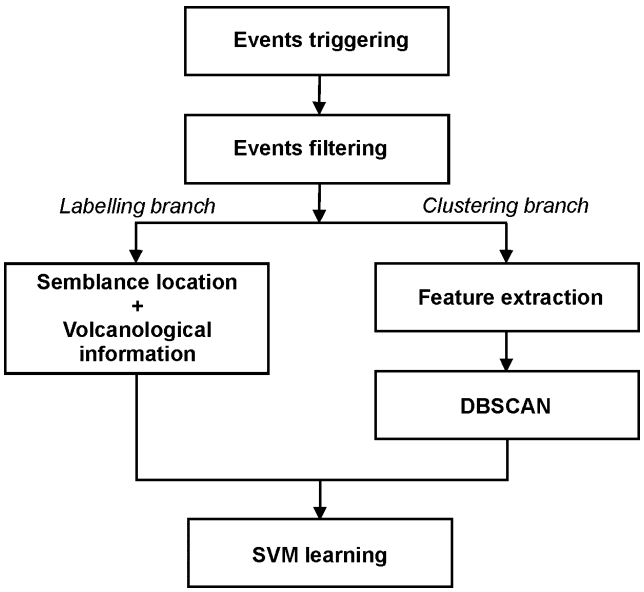

Figure 9. Scheme of the learning system.

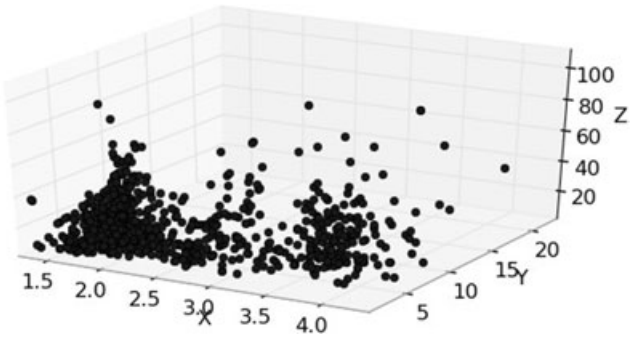

Figure 10. Feature space with frequency, quality factor and peak-to-peak amplitude of the infrasound events recorded at EBEL station during 2007 September–November 2007.

Such last two clusters were related to different kinds of explosive activity at SEC. In particular, the events belonging to cluster 3 were coincident with 'more visible' explosions, characterized by a relevant presence of ash, whereas the events of cluster 2 were hardly visible in the monitoring video-camera recordings (Cannata *et al.*

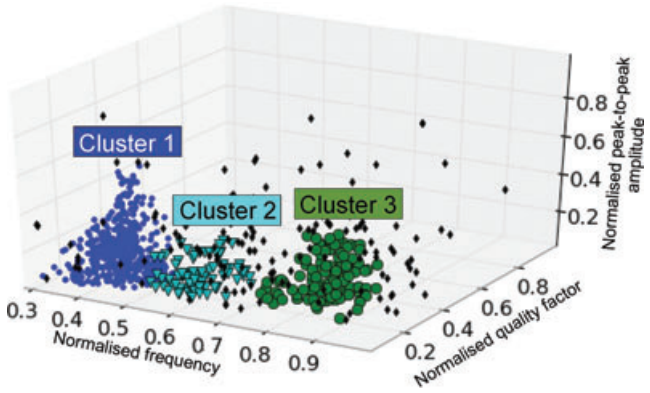

Figure 11. Clustering of the feature space reported in Fig. 10. The clusters are indicated with blue (cluster 1) and green circles (cluster 3) and light green triangles (cluster 2), the outliers with black diamonds.

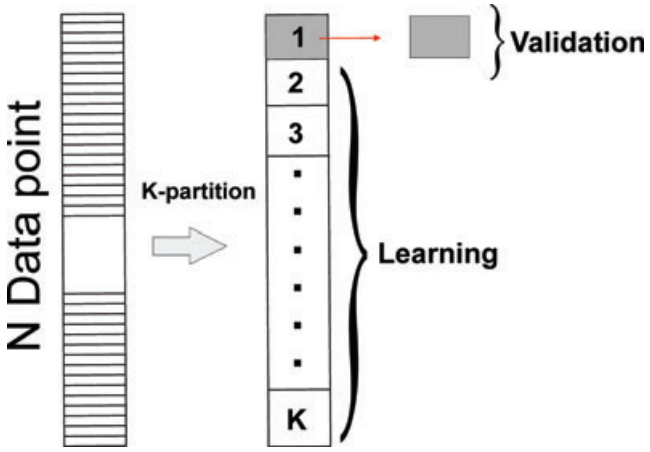

Figure 12. Basic scheme of K-Fold Cross-validation (see Section 4.3 for details).

2009b). Features clustering together with labels provide the patterns for SVM learning process.

As mentioned in Section 4.2, optimization of parameters C (regularization parameter) and σ (radial basis function kernel parameter) is a key step in SVM learning because their values determine classification performance (Devos *et al.* 2009). As a consequence, model selection is applied with the aim of finding the best pair of parameters C and σ that minimizes the error rate estimated as the ratio between misclassified and hit patterns. These parameters can be chosen using a cross-validation (CV) approach (Hastie *et al.* 2002), which is a statistical method for learning algorithms evaluation and model selection. In particular, in K-fold CV the available data set is partitioned into K subsets or 'folds': K–1 folds are used for SVM learning purpose, and the remaining fold for model validation (Fig. 12). Thus, K iteration of learning and validation are performed and for each ith iteration the training process is carried out using $K-1$ folds and the ith fold for validation (Fig. 12). All SVM training algorithms are computed using one-against-all method (see Section 4.2). Since we worked on a small data set, a simple exhaustive grid search can be performed (Hsu *et al.* 2007). In particular, C was systematically changed in the range [1–100] with a step of 10, σ in the range [0.1–10] with a step of 0.5 and a K-fold CV with $K = 10$ was used. The entire procedure can be summarized as follows (Fig. 13): (1) a grid value of C and σ is defined; (2) for each pair of C and σ values, a mean error rate is computed averaging the error rate values obtained by the K SVM models; (3) the pair of C and σ with the minimum error rate is selected; (4) such a pair is used to train the final SVM model with the whole data set, comprising all the K folds. Here, the best parameter values were $C = 1$ and $\sigma = 0.1$, for which mean CV error minimized to 0.6 per cent.

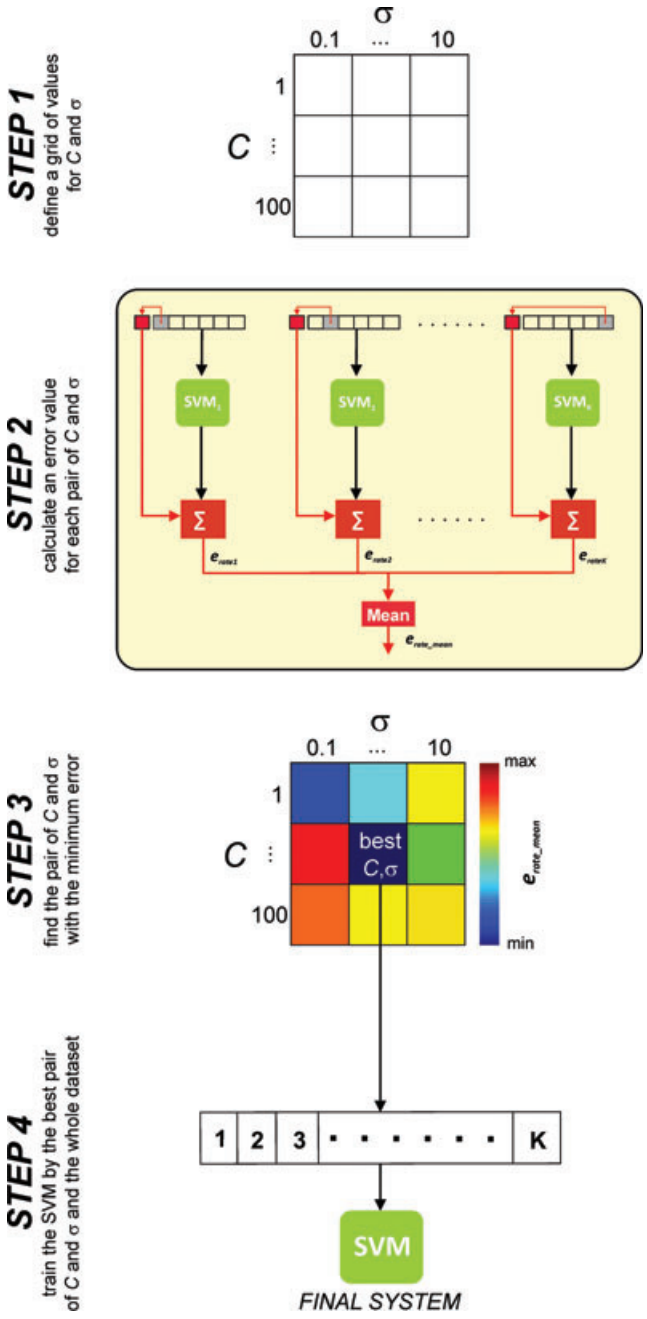

Figure 13. Best SVM model selection using K-Fold Cross-validation (see Section 4.3 for details).

4.4 Testing phase and final system

To verify the system, the trained SVM is tested by classifying new unknown infrasonic events and then assigning them to their source crater. The reliability is verified using events not analysed during the previous learning phase (Section 4.3). To this end, a new test data set of about 610 events, recorded during 2 months, 2007 August and December, was used and labelled by location algorithm based on

Table 1. Confusion matrix calculated in the testing phase. Each column represents the instances in the predicted class (based on the SVM model), while each row represents the instances in the actual class (based on the previously attributed labels). Thus, the entries on the diagonal (bold numbers) count the events in which prediction agrees with known labels, whereas the other entries the misclassified events.

		Predicted		
		Cluster 1	Cluster 2	Cluster 3
ACTUAL	Cluster 1	**476**	9	6
	Cluster 2	9	**15**	8
	Cluster 3	8	33	**46**

semblance method (Section 3.3). Moreover, the events belonging to cluster 2 and cluster 3 were labelled using information related to the intensity of the explosive activity (Cannata *et al.* 2009b). The quality of classification is quantified using confusion matrix (Table 1), where each column represents the instances in the predicted class (based on the SVM model), while each row represents the instances in the actual class (based on the previously attributed labels). Thus, the entries on the diagonal count the events in which prediction agrees with known labels, whereas the other entries the misclassified events. 63 elements were wrong assigned, providing an error rate of about 11.97 per cent. Misclassifications were mostly concentrated in the second and third classes that are related to the two different explosion activities of SEC crater. Indeed, such a distinction is qualitative and not clearcut, hence many halfway events can be misclassified. If we do not take into account the distinction between clusters 2 and 3, and consider them as a single cluster, the error decreases to 5.25 per cent.

Finally, the proposed system can be summarized as follows (Fig. 14): (i) triggering procedures is performed on buffer of acquired signal; (ii) then, if events are found, the system evaluates whether there is a sufficient number of stations for semblance location algorithm; (iii) if the number of stations is not sufficient, alternative 'single station' location is performed by extracting signal features and classifying them using the trained SVM. It is also worth noting that SVM classifier is also applied offline on localizable events to evaluate its performance in distinguishing NEC events (cluster 1) from SEC events (clusters 2 and 3). In this case, events belonging to clusters 2 and 3 are simply considered SEC events and then labelled based on the source vent, with no further distinction depending on the type of explosive activity. This task is carried out by comparing the results of the classifier with the location parameters provided by the semblance algorithm. By the inspection of the obtained error rate, a new clustering execution is necessary when classification of new signals is not aligned with that of infrasonic network classifier. This may be caused by the creation of a new active vent or by the changing activity of a pre-existing vent; in such a case the system must be updated.

5 CONCLUSIONS

At active volcanoes the detection and location of explosive activity is generally obtained by videocameras and thermal sensors (Harris *et al.* 1997; Bertagnini *et al.* 1999). However, the efficiency of such instruments is severely reduced or inhibited in case of poor visibility caused by clouds or gas plumes. In these cases, the detection and characterization of explosive activity by infrasound is very useful (e.g. Cannata *et al.* 2009a) and some techniques, based on infrasound signals recorded by arrays or networks, were developed to locate the source of this signal and therefore the active vent (e.g. Ripepe & Marchetti 2002). All these techniques require that most of the stations work properly and that the noise level is low. Unfortunately, sometimes during winter season, because of bad weather conditions, the possible lack of signals from some summit stations prevents applying the aforementioned standard location algorithms. At Mt. Etna, the events at a single vent for a certain type of

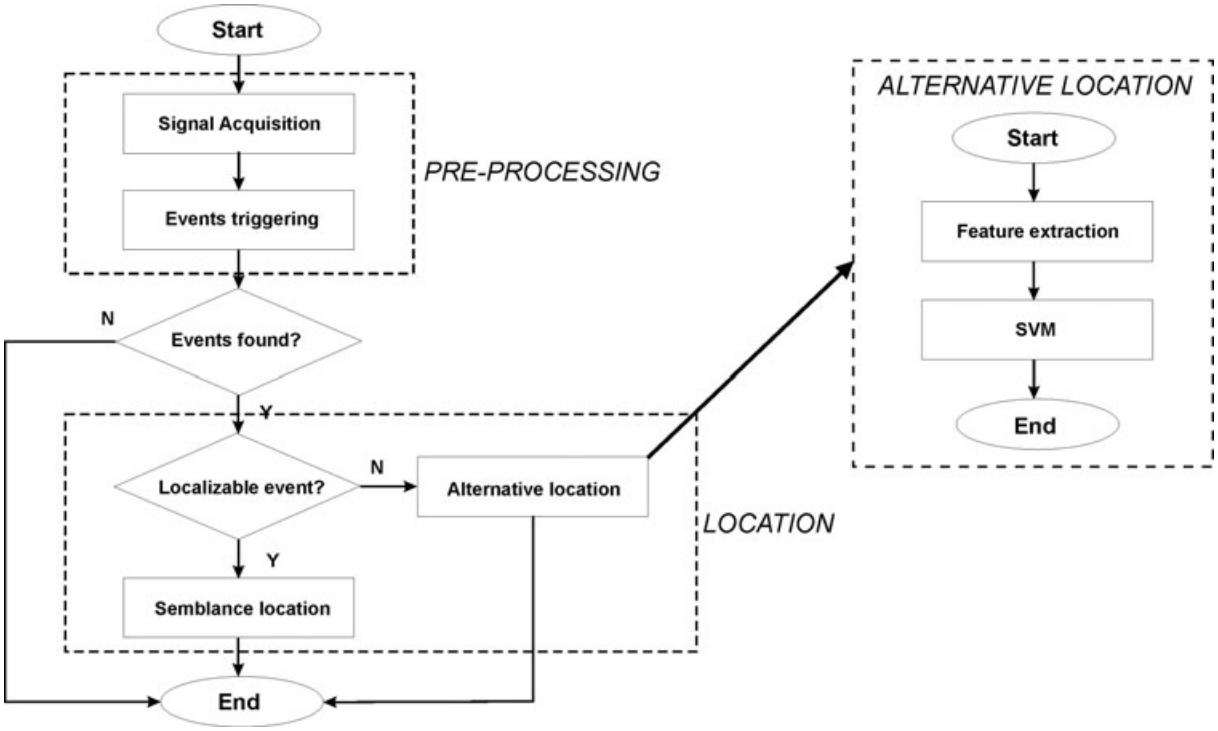

Figure 14. Flow chart of the proposed location system.

activity maintain their features stable over time (Cannata *et al.* 2009b). Therefore, once the link between event characteristics and vent is known we can understand which crater is active and which volcanic activity is going on by simply extracting the features of the infrasonic signal at a single station (dominant frequency, quality factor and peak-to-peak amplitude). In the light of this, a system, whose learning phase is based on clustering (DBSCAN) and classification techniques (SVM), together with geophysical information, was developed. After the training phase this system automatically allows recognizing the active vent, with no location algorithm and by using only a single station with a success of ∼95 per cent.

It should be noted, however, that in a volcano as Etna, characterized by almost continuous eruptive activity with different styles and topographical variations of the summit area, the vent geometry can change and consequently also the infrasound spectral features. Therefore, spectral characterization and source location must be considered complementary, especially when long lasting periods are investigated.

ACKNOWLEDGMENTS

We are grateful to the Editor Matthias Hort and two anonymous reviewers for their constructive criticism. We thank Stephen Conway for revising and improving the English text. Work partly performed with grants of the 'Flank project' (INGV-DPC 2007–2009).

REFERENCES

Behncke, B. & Neri, M., 2003. Cycles and trends in the recent eruptive behaviour of Mount Etna (Italy), *Can. J. Earth Sci.*, **40**, 1405–1411.

Berkhin, P., 2002. *Survey Of Clustering Data Mining Techniques*, Accrue Software, San Jose, CA.

Bertagnini, A., Coltelli, M., Landi, P., Pompilio, M. & Rosi, M., 1999. Violent explosions yield new insights into dynamics of Stromboli volcano, *EOS, Trans. Am. geophys. Un.*, **80**, 633–636.

Burges, C.J.C., 1998. A tutorial on support vector machines for pattern recognition, *Data Min. Knowl. Discov.*, **2**, 121–167.

Cannata, A., Catania, A., Alparone, S. & Gresta, S., 2008. Volcanic tremor at Mt. Etna: inferences on magma dynamics during effusive and explosive activity, *J. Volc. Geotherm. Res.*, **178**, doi:10.1016/j.jvolgeores.2007.11.027.

Cannata, A., Montalto, P., Privitera, E., Russo, G. & Gresta, S., 2009a. Tracking eruptive phenomena by infrasound: May 13, 2008 eruption at Mt. Etna, *Geophys. Res. Lett.*, **36**, doi:10.1029/2008GL036738.

Cannata, A., Montalto, P., Privitera, E. & Russo, G., 2009b. Characterization and location of infrasonic sources in active volcanoes: Mt. Etna, September–November 2007, *J. geophys. Res.*, **114**, doi:10.1029/2008JB006007.

Cheng, J., Randall, A. & Baldi, P., 2006. Prediction of protein stability changes for single-site mutations using support vector machines, *Proteins*, **62**, 1125–1132, doi:10.1002/prot.20810.

Davies, D.L. & Bouldin, D.W., 1979. A cluster separation measure, *IEEE Trans. Pattern Anal. Mach. Intell.*, **1**, 224–227.

Devos, O., Ruckebusch, C., Durand, A., Duponchel, L. & Huvenne, J.P., 2009. Support vector machines (SVM) in near infrared (NIR) spectroscopy: focus on parameters optimization and model interpretation, *Phys. Chem. B*, **113**, 6031–6040.

Di Grazia, G., Cannata, A., Montalto, P., Patanè, D., Privitera, E., Zuccarello, L. & Boschi, E., 2009. A new approach to volcano monitoring based on 4D analyses of seismo-volcanic and acoustic signals: the 2008 Mt. Etna eruption, *Geophys. Res. Lett.*, **36**, doi:10.1029/2009GL039567.

Ester, M., Kriegel, H.P., Sander, J. & Xu, X., 1996. A density-based algorithm for discovering clusters in large spatial databases with noise, *Proc. KDD,* **96,** 226–231.

Firstov, P.P. & Kravchenko, N.M., 1996. Estimation of the amount of explosive gas released in volcanic eruptions using air waves, *Volcanol. Seismol.*, **17**, 547–560.

Fukao, Y. & Suda, N., 1989. Core modes of the Earth's free oscillations and structure of the inner core, *Geophys. Res. Lett.*, **16**, 401–404.

Garces, M., Harris, A., Hetzer, C., Johnson, J., Rowland, S., Marchetti, E. & Okubo, P., 2003. Infrasonic tremor observed at Kilauea Volcano, Hawaii, *Geophys. Res. Lett.*, **30**, 2023, doi:10.1029/2003GL018038.

Gresta, S., Ripepe, M., Marchetti, E., D'Amico, S., Coltelli, M., Harris, A.J.L. & Privitera, E., 2004. Seismoacoustic measurements during the July–August 2001 eruption at Mt. Etna volcano, Italy, *J. Volc. Geotherm. Res.*, **137**, 219–230.

Hagerty, M.T., Schwartz, S.Y., Garces, M.A. & Protti, M., 2000. Analysis of seismic and acoustic observations at Arenal Volcano, Costa Rica, 1995–1997, *J. Volc.Geotherm. Res.*, **101**, 27–65.

Harris, A.J.L., Blake, S., Rothery, D.A. & Stevens, N.F., 1997. A chronology of the 1991 to 1993 Mount Etna eruption using advanced very high resolution radiometer data: implications for real-time thermal volcano monitoring, *J. geophys. Res.*, **102**, 7985–8003.

Hastie, T., Tibshirani, R. & Friedman, J., 2002. *The Elements of Statistical Learning*, p. 533, Springer, New York.

Hori, S., Fukao, Y., Kumazawa, M., Furumotom M. & Yamamoto, A., 1989. A new method of spectral analysis and its application to the Earth's free oscillations: the "Sompi" method, *J. geophys. Res.*, **94**(B6), 7535–7553.

Hsu, C.W., Chang C.C. & Lin, C.J., 2007. A practical guide to support vector classification, http://www.csie.ntu.edu.tw/∼cjlin/papers/guide/guide.pdf (last accessed 2011 January 31).

Hsu, C.W. & Lin, C.J., 2002. A comparison of methods for multi-class support vector machines, *IEEE Trans. Neural Netw.*, **13**, 415–425.

Hwanjo, Y., Yang, J. & Han, J., 2003. Classifying large data sets using SVMs with hierarchical clusters, *in Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Washington, DC.

Johnson, J.B., 2005. Source location variability and volcanic vent mapping with a small-aperture infrasound array at Stromboli Volcano, Italy, *Bull. Volcanol.*, **67**, 1–14.

Johnson, J.B. & Lees, J.M., 2000. Plugs and chugs: seismic and acoustic observations of degassing explosions at Karymsky, Russia and Sangay, Ecuador, *J. Volc. Geotherm. Res.*, **101**, 67–82.

Johnson, J.B., Lees, J. & Varley, N., 2010. Characterizing complex eruptive activity at Santiaguito, Guatemala using infrasound semblance in networked arrays, *J. Volc. Geotherm. Res.*, **199**, doi:10.1016/j.jvolgeores.2010.08.005.

Jones, K.R., Johnson, J., Aster, R., Kyle, P.R. & McIntosh, W.C., 2008. Infrasonic tracking of large bubble bursts and ash venting at Erebus volcano, Antarctica, *J. Volc. Geotherm. Res.*, **177**, doi:10.1016/j.jvolgeores.2008.02.001.

Joswig, M., 1990. Pattern recognition for earthquake detection, *Bull. seism. Soc. Am.*, **80**(1), 170–186.

Kay, S.M., 1988. *Modern Spectral Estimation, Theory and Application*, Prentice-Hall, Englewood Cliffs, NJ.

Kecman, V., 2001. *Learning and Soft Computing. Support Vector Machines, Neural Networks, and Fuzzy Logic Models*, The MIT Press, Cambridge.

Kohler, A., Ohrnberger, M. & Scherbaum, M., 2009. Unsupervised feature selection and general pattern discovery using Self-Organizing Maps for gaining insights into the nature of seismic wavefields, *Comput. Geosci.*, **35**, 1757–1767, doi:10.1016/j.cageo.2009.02.004.

Kumazawa, M., Imanishi, Y., Fukao, Y., Furumoto, M. & Yamamoto, A., 1990. A theory of spectral analysis based on the characteristic property of a linear dynamic system, *Geophys. J. Int.*, **101**, 613–630.

Langer, H., Falsaperla, S., Masotti, M., Campanini, R., Spampinato, S. & Messina, A., 2009. Synopsis of supervised and unsupervised pattern classification techniques applied to volcanic tremor data at Mt Etna, Italy, *Geophys. J. Int.*, **178**, doi:10.1111/j.1365–246X.2009.04179.x.

Lesage, P., 2008. Automatic estimation of optimal autoregressive filters for the analysis of volcanic seismic activity, *Nat. Hazards Earth Syst. Sci.*, **8**, 369–376.

Marchetti, E., Ripepe, M., Ulivieri, G., Caffo, S. & Privitera, E., 2009. Infrasonic evidences for branched conduit dynamics at Mt. Etna volcano, Italy, *Geophys. Res. Lett.*, **36**, L19308, doi:10.1029/2009GL040070.

Marple, S.L., 1987. *Digital Spectral Analysis with Applications*, Prentice Hall, Englewood Cliffs.

Mars, J., Lacoume, J.L., Mari, J.L. & Glangeaud, F., 2004. *Traitement du Signal Pour Géologues et géophysiciens*, Vol. 3, Techniques avancées, Technip.

Masotti, M., Campanini, R., Mazzacurati, L., Falsaperla, S., Langer, H. & Spampinato, S., 2008. TREMOrEC: a software utility for automatic classification of volcanic tremor, *Geochem. Geophys. Geosyst.*, **9**(1), Q04007, doi:10.1029/2007GC001860.

Matoza, R.S., Hedlin, M. & Garces, M., 2007. An infrasound array study of Mount St. Helens, *J. Volc. Geotherm. Res.*, **160**, 249–262.

Montalto, P., Cannata, A., Privitera, E., Gresta, S., Nunnari, G. & Patanè, D., 2010. Towards an automatic monitoring system for infrasonic events at Mt. Etna: strategies for source location and modelling, *Pure appl. Geophys.*, **167**, doi:10.1007/s00024–010-0051-y.

Neidell, N. & Taner, M.T., 1971. Semblance and other coherency measures for multichannel data. *Geophysics*, **36**, 482–497, doi:10.1190/1.1440186.

Noble, W.S., 2004. Support vector machine applications in computational biology, in *Kernel Methods in Computational Biology*, pp. 71–92, eds Schölkopf, B., Tsuda, K., Vert, J., The MIT Press, Cambridge, MA.

Ohrnberger, M., 2001. Continuous automatic classification of seismic signals of volcanic origin at Mt. Merapi, Java, Indonesia, *PhD thesis*, University of Potsdam, 158 pp.

Ripepe, M. & Marchetti, E., 2002. Array tracking of infrasonic sources at Stromboli volcano, *Geophys. Res. Lett.*, **29**(22), 2076, doi:10.1029/2002GL015452.

Ripepe, M., Poggi, P., Braun, T. & Gordeev, E., 1996. Infrasonic waves and volcanic tremor at Stromboli, *Geophys. Res. Lett.*, **23**, 181–184.

Rowe, C.A., Aster, R.C., Kyle, P.R., Dibble, R.R. & Schlue, J.W., 2000. Seismic and acoustic observations at Mount Erebus Volcano, Ross Island, Antarctica, 1994–1998, *J. Volc. Geotherm. Res.*, **101**, 105–128.

Ruiz, M.C., Lees, J.M. & Johnson, J.B., 2006. Source constraints of Tungurahua volcano explosion events, *Bull. Volcanol.*, **68**, 480–490.

Vapnik, V.N., 1998. *Statistical Learning Theory*, John Wiley & Sons, New York.

Vergniolle, S. & Brandeis, G., 1994. Origin of the sound generated by Strombolian explosions, *Geophys. Res. Lett.*, **21**, 1959–1962.

Weston, J. & Watkins, C., 1999. *Multi-class support vector machines*, presented at the *Proceedings ESANN99*, ed. M. Verleysen, Brussels, Belgium.

Withers, M., 1997. An automated local/regional seismic event detection and location system using waveform correlation, *PhD thesis*, New Mexico Tech., Socorro, NM.

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

APPENDIX: SOMPI METHOD

Time-series modelling consists of estimating the governing dynamics of the hypothetical linear system that has yielded the given time-series data (Kumazawa *et al.* 1990). In these approaches, a signal is considered as the impulse response of an AR or an autoregressive moving average (ARMA) filter. In general, ARMA filter is a discrete-time system that takes an input sequence x_n and produces an output sequence y_n. This kind of system can be described by a linear-constant difference equation

$$x_n = \sum_{k=1}^{p} a_k x_{n-k} - \sum_{k=1}^{q} b_k y_{n-k}, \tag{A1}$$

where $\{a_k\}$ and $\{b_k\}$ are the system coefficients, p and q are the order of the AR and MA parts of the filter, respectively. The coefficients of the AR filter can be obtained by solving the modified Yule–Walker equation (Marple 1987) and the coefficients of the MA filter can be estimated using the Durbin method (Kay 1988; Mars *et al.* 2004). As argued in Lesage (2008), this process is affected by numerical instabilities and long computation time. Furthermore, the deconvolution of the AR part alone gives good estimation of the duration and spectral content of the considered signals (Lesage 2008). To estimate the AR coefficients, the Sompi method (Kumazawa *et al.* 1990) can be implemented. Unlike the traditional spectral estimators in real frequency space, this method yields a line-shaped spectrum in complex frequency space. The basic concepts of the AR model and the formulation based on the maximum likelihood principle lead to a model estimation algorithm different from other AR methods (Fukao and Suda 1989; Kumazawa *et al.* 1990). By Sompi analysis, a time-series is deconvoluted into a linear combination of coherent oscillation with decaying amplitude and additional noise. Let (x_n) time-series that can be considered the sum of signal (u_n) and Gaussian white noise (e_n)

$$x_n = u_n + e_n, \tag{A2}$$

where u_n is described as a set of decaying sinusoids

$$u_n = \sum_k \{C_k(z_k)^n + C_k^*(z_k^*)^n\} \tag{A3}$$

and z_k is defined as

$$z_k = \exp(2\pi(g_k + \mathrm{i}\, f_k)\Delta t), \tag{A4}$$

where Δt is the sampling step and the symbol $*$ represents the complex conjugate. In eq. (A3) C_k represents the complex amplitude of the kth sinusoid at the complex frequency given by $f_k - \mathrm{i}g_k$ and i is $\sqrt{-1}$. The time-series (u_i) is defined as the sequence satisfying the AR equation

$$\sum_{j=-m}^{m} a_j u_{i-j} = 0, \tag{A5}$$

where $(a_j; j = -m, \ldots, m)$ are real AR coefficients. An exhaustive treatment about a_j coefficients estimations is reported in Hori *et al.* (1989), Fukao and Suda (1989) and Kumazawa *et al.* (1990). Briefly, a way to compute the coefficients a_j that satisfy eq. (A5) is the minimization of the functional S

$$S = \sum_{i=-N+m}^{N-m} \left(\sum_{j=-m}^{m} a_j x_{i-j} \right)^2 \tag{A6}$$

under the condition

$$\sum_{j=-m}^{m} a_j^2 = 1. \tag{A7}$$

This minimization problem leads to an eigenvalue problem where coefficients a_j are the eigenvectors corresponding to minimum eigenvalues. Now, once the a_j are calculated, the Sompi characteristic equation is defined as

$$\sum_{j=-m}^{m} a_j z^{-j} = 0. \tag{A8}$$

The roots z_k and z_k^* of eq. (A8) give the complex frequencies expressed in eq. (A4). Let (x_i) a time-series, Sompi method extracts m wave elements characterized by a complex frequency $f_k - ig_k$ where f_k is the frequency, g_k is the growth rate.