Journal of Volcanology and Geothermal Research 282 (2014) 134–147

Contents lists available at ScienceDirect

Journal of Volcanology and Geothermal Research

journal homepage: www.elsevier.com/locate/jvolgeores

Pattern recognition applied to seismic signals of the Llaima volcano (Chile): An analysis of the events' features

Millaray Curilem [a,*], Jorge Vergara [e], Cesar San Martin [a], Gustavo Fuentealba [b], Carlos Cardona [c], Fernando Huenupan [a], Max Chacón [d], M. Salman Khan [e], Walid Hussein [e], Nestor Becerra Yoma [e]

[a] Department of Electrical Engineering, Universidad de La Frontera, Av. Francisco Salazar 01145. Temuco, Chile
[b] Department of Physics Universidad de La Frontera, Av. Francisco Salazar 01145. Temuco, Chile.
[c] Observatorio Vulcanológico de los Andes Sur, Rudecindo Ortega 03850, Temuco, Chile
[d] Department of Informatics Engineering, Universidad de Santiago de Chile, Avenida Ecuador #3659. Estación Central, Santiago, Chile
[e] Department of Electrical Engineering, Universidad de Chile, Av. Tupper 2007, P.O. Box 412-3, Santiago, Chile

ARTICLE INFO

Article history:
Received 3 December 2013
Accepted 12 June 2014
Available online 21 June 2014

Keywords:
Seismic discrimination
Volcano monitoring
Signal processing
Pattern recognition
Support vector machines

ABSTRACT

This paper proposes a computer-based classifier to automatically identify four seismic event classes of the Llaima volcano, one of the most active volcanoes in the Southern Andes, situated in the Araucanía Region of Chile. A combination of features that provided good recognition performance in our previous papers concerning the Llaima and Villarica (located 100 km south of Llaima) volcanoes is utilized in order to train the classifiers. These features are extracted from the amplitude, frequency and phase of the seismic signals. Unlike the previous works where fixed length windows were used to obtain the seismic signals, this paper employs signals of variable lengths that span the entire seismic event. The classifiers are implemented using support vector machines. A confidence analysis is also included to improve reliability of the classification. Results indicate that the features used for recognition of the events of Villarica volcano also provide good recognition results for the Llaima volcano, yielding classification exactitude of over 80%.

1. Introduction

The importance of developing tools for the automatic detection of volcanic events is due to the increasing need to monitor active volcanoes in order to understand their internal dynamics, model their behavior and establish criteria to act efficiently in the case of an eruption (Cannata et al., 2013). Moreover, the monitoring should be done around the clock all year long. Volcanic event detection requires specialized domain knowledge. Since each volcano exhibits a particular behavior, the analysis should be individualized.

For its geographical location, Chile has a high volcanic activity. There are about two thousand volcanoes, of which around one hundred have been considered historically active. The volcanic chain in the Southern Andes is part of an active tectonic boundary between the Nazca oceanic plate and the continental plate of South America at latitude 38.4° S is Llaima, located in the Araucanía Region (38° 41′ S–71° 44′ W), on the western edge of the Andes, as shown in Fig. 1. It is a complex active strato-volcano, with a mainly andesitic–basaltic composition and is 3125 m high, rising some 1200 m above the surrounding summits. The total height of the volcano is estimated to be 2400 m above its base with irregular topography. The main building consists of two peaks, the most prominent being the North one (3125 m), which is separated by 1 km from the South Summit pass or Pichillaima (2920 m). The tallest summit has an open crater of about 350 m in diameter with remarkably active fumaroles, which can be spotted from great distances. One of the latest eruptive cycles in Llaima began in May 2007 with a weak ash emission followed by a moderate Strombolian eruption and lahar generation in January 2008, which culminated in April 2009 with a vigorous Strombolian eruption.

The Southern Andes Volcano Observatory (OVDAS) is the state agency responsible for establishing systems to continuously monitor and record over forty active Chilean volcanoes. The monitoring focuses mainly on seismic data, but also incorporates other measurements, including deformation and geochemistry.

Volcanic seismicity has a prominent role in monitoring volcanoes (Zobin, 2012) because it provides useful information, that propagates over long distances which allows to process remotely in almost real time the volcanic activity. Signals are measured by seismic stations located at different parts of the volcano's external structure. Today Llaima is monitored by twelve stations. The seismic events suggest internal processes occurring inside the volcano's structure. OVDAS uses the

* Corresponding author.
E-mail addresses: millaray.curilem@ufrontera.cl (M. Curilem), cesar.sanmartin@ufrontera.cl (C.S. Martin), gustavo.fuentealba@ufrontera.cl (G. Fuentealba), carlos.cardona@sernageomin.cl (C. Cardona), fernando.huenupan@ufrontera.cl (F. Huenupan), max.chacon@usach.cl (M. Chacón), mkhan@ing.uchile.cl (M. Salman Khan), whussein@ing.uchile.cl (W. Hussein), nbecerra@ing.uchile.cl (N.B. Yoma), jorge.vergara.q@gmail.com (J. Vergara).

http://dx.doi.org/10.1016/j.jvolgeores.2014.06.004
0377-0273/

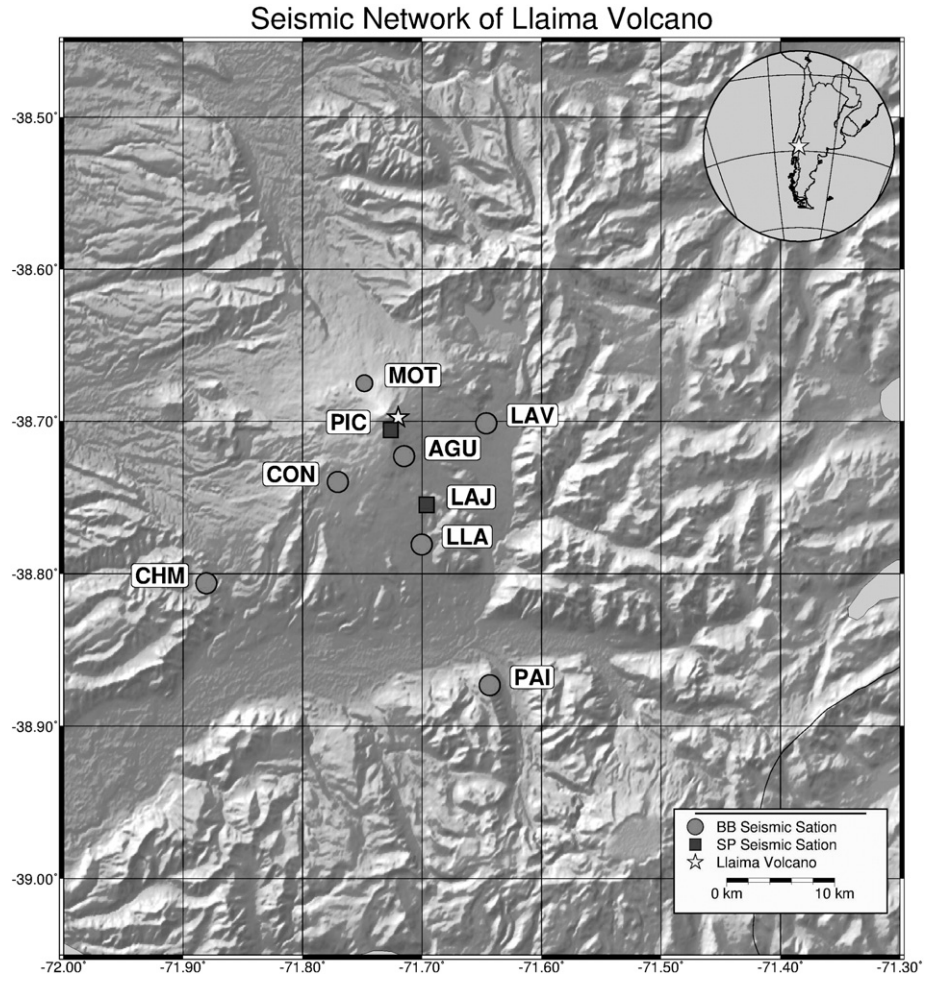

Fig. 1. Location of Llaima and its stations.

criteria defined in Lahr et al. (1994) to identify the three volcanic events considered in this work. One of the major seismic events is the tremor (TR), the genesis of which is associated with degassing, fluctuations in gaseous phases, and resonance processes in the internal conduits (McNutt, 1992; Julian, 1994). Another important event class is the long-period (LP) event, which is related to the pressure of gas and other fluids in the conduit, but which happens at discrete periods (Chouet, 1992). The volcano-tectonic (VT) event is associated to the fracture of solid parts of the volcano or the conduits. The automatic detection of these events requires sophisticated pattern recognition schemes because their characteristics are highly dynamic and may lead to disagreement between experienced analysts. This problem has not yet been resolved and is the driving force behind a lot of research in this area.

The challenge of monitoring volcanoes includes the need to incorporate tools that automate the identification of volcanic activity. To do this, tools from the area of signal processing and pattern recognition are utilized. Within signal processing, the problem is generally tackled in two stages: feature extraction and classification. The feature extraction stage defines what information (i.e. features) is extracted from the signal to facilitate discrimination between different types of events. In the classification stage the design and implementation of the classifiers are performed.

In Scarpetta et al. (2005), a linear predictive coding technique was proposed to extract spectral features, and parameterization of the signal to extract information about the waveform. In Langer et al. (2006), autocorrelation functions, obtained by fast Fourier transform (FFT), were used to represent the spectral content. An amplitude ratio was applied to distinguish between signal peaks and long duration with similar frequency content. Other widely used methods are the wavelet transform (Dowla, 1995; Gendron and Nandram, 2001; Erlebacher and Yuen, 2004), and cross-correlation methods (Lesage et al., 2002). In Ibáñez et al. (2009) the authors worked with signals from the Stromboli and Etna volcanoes in Italy. Hidden Markov models (HMMs) were used to model and identify different online patterns using spectrum features, as in Beyreuther et al. (2008) where a system based on HMMs was employed to detect and classify volcanic seismicity and/or tectonic earthquakes from noise, in continuous seismic data, recorded on the volcanic island of Tenerife. A similar approach has been used in a simpler form as discrete HMM (Ohrnberger, 2001). Other techniques for pre-processing of these signals consider methods of independent component analysis and information theory, as presented in Amari and Cichocki (1998). In Messina and Langer (2011), short-time Fourier transform (STFT) was applied to identify changes in the activity of Etna tremor. The researchers used an unsupervised classification method for the early and reliable identification of changes in the tremor.

The machine learning approaches, like neural networks and support vector machines, have the advantage of allowing automatic pattern recognition independent of phenomenological knowledge of the volcanic processes. This is why these techniques are widely used in the analysis and classification of volcanic seismic data. The most common procedure for an identification system is based on supervised classification. As in Falsaperla et al. (1996) that describe a system with a multilayer perceptron for the classification of 'explosion quakes' at the Stromboli volcano to identify four different classes of shocks. The proposed scheme correctly classified 89% of the events, demonstrating that the

automatic technique is reliable, encouraging further applications in the field of volcanic seismology. In Joevivek et al. (2010), the authors used amplitude statistics (mean, standard deviation, skewness and kurtosis), and incorporated a statistical wavelet decomposition phase and power to detect seismic events in southern India. They proposed a system to detect small earthquakes, distinguishing between seismic and non-seismic sources. They compared various types of classifiers and reported that a support vector machine (SVM) provided the best performance with an accuracy of 94%. This conclusion supported the results of Giacco et al. (2009) and Langer et al. (2009), who also achieved the best results with a SVM classifier.

Several authors use unsupervised methods, like self-organizing maps (SOMs), to cluster volcanic events with similar behavior. Carniel et al. (2013a) propose a technique to improve data analysis and highlight possible dynamics or precursory regimes by employing SOMs. The authors demonstrated a practical application on the data recorded in Raoul Island around the March 2006 phreatic eruption, which revealed both a diurnal anthropogenic signal and post-eruption system excitation. A similar approach was proposed in Carniel et al. (2013b) where the authors applied SOM to assess the low-level seismic activity prior to small-scale phreatic events in the Ruapehu volcano in New Zealand. In Esposito et al. (2008) a method based on an unsupervised neural network was presented to cluster the waveforms of very-long-period (VLP) events associated with explosive activity at the Stromboli volcano. They applied this method to investigate the relationship between each event and its associated VLP explosive waveform. In Cannata et al. (2011) features that characterize the infrasound events were extracted to define three clusters. Volcanic information concerning the intensity of the explosive activity was associated with each cluster, and with a particular source of event and/or a kind of volcanic activity. A SVM classifier was employed to maximize the margins of separation among the clusters. Using only a single station this scheme provided a 95% accuracy. In Langer et al. (2009) a comparative study was presented to compare two supervised methods, SVM and a multilayer perceptron neural network (MLP), and two unsupervised techniques, cluster analysis and self-organizing maps. The authors dealt with tremor signals from Etna, defining four classes: pre-eruptive, eruptive, post-eruptive events and lava fountains. The SVM classifier gave superior results (>90%) compared to MLP (>80%). The clustering methods enabled the observation that the characteristics of the tremor changed over time.

Volcanoes of the Araucanía Region of Chile have also been studied, in particular the Villarrica and Llaima volcanoes (Vila et al., 2006; Mora-Stock et al., 2012). Their structure, composition and volcanic building are similar. Seismic activity of Villarrica volcano is characterized by an active lava lake in the summit crater, with mild Strombolian style eruptions of basaltic–andesitic products (Richardson and Waite, 2013). Three types of seismic events are present: LP, TR and VT. LP events are very shallow signals with no clear S phase. They are signals with low frequencies in a range of 1 to 5 Hz. Harmonic tremor corresponds to a continuous stream of vibrations on the seismic record, also characterized by frequencies in the 1 to 5 Hz range. The VT earthquakes have depths ranging from 1 to 20 km. Their P and S phases are well-defined and they have a wide range of similar tectonic earthquake frequencies. Seismic activity of Llaima volcano is characterized in this work by a post-eruptive stage, related to an open conduit volcanic phase. The seismic signals were related with fluid movement inside the volcanic conduits, sometimes correlated with surface changes of the ash and gas plumes. The volcano-related seismic activity was characterized by LP events, with frequencies in the range of 1.0–1.5 Hz and VT earthquakes with magnitudes generally less than or equal to 3.0 M_L (local magnitude). Tremor episodes were also recorded ranging from 1.0 Hz to 1.3 Hz. In both volcanoes tremors could be formed by a coalescing sequence of transients mainly due to degassing through the ducts; agreeing with the description suggested by Chouet (1992) and Julian (1994). While in the Llaima volcano this is not a permanent signal, in the Villarrica tremor is always present, since the ducts are open.

In Curilem et al. (2009), 1033 signals of Villarrica were considered, including events of types LP, TR and tremor energy (TE). The TE event is a specific type of tremor that represents the increase of amplitude due to the obstruction of conduits. A set of other events (OT) was also considered, composed of background noise, tectonic earthquakes and all other signals which do not belong to the LP, TR and TE groups. The authors segmented the signal into 30-second frames, and implemented an artificial neural network to classify the patterns. The feature selection process showed that the mean, median, maximum amplitude, energy in the frequency band of 1.56–3.12 Hz and peak frequency were good signal descriptors. The classifier achieved a Kappa coefficient of $\kappa = 0.89 \pm 0.04$ for the classification of the four events (i.e. four classes). The Kappa coefficient is a measure of agreement between results of the numerical classifier and the classification by an expert.

In San-Martin et al. (2010), 893 signals were used to classify different events of Llaima. The events considered were LP, OT and VT, and were determined by experts on segmented windows of 1 min each. This work applied circular statistics, which is a measure that provides insight into the behavior of the phase of signal typically obtained using a Hilbert transform. The instantaneous phase of the signal was considered a circular random variable in the range of $[0, 2\pi)$. More specifically, the first circular moment was used together with the wavelet energy in the same sub-band than the work presented in Curilem et al. (2009). A linear discriminator allowed a 92.54% correct classification rate of the three classes.

It can be observed from the above bibliographical review that a fundamental issue of pattern recognition is feature extraction (Álvarez et al., 2012). Feature extraction defines what information can be extracted from the signals for discrimination. It is one of the most important problems to be addressed in the context of volcano event classification.

The contribution of this paper is as follows. Since Villarrica and Llaima volcanoes have similar structure and composition, in the current paper the five features proposed in Curilem et al. (2009), to classify events from Villarrica, are also considered to classify the Llaima events. A sixth feature proposed in San-Martin et al. (2010), i.e. the circular statistics of the phase, is also included. All the features from Curilem et al. (2009) considered here were chosen because they led to a high classification accuracy. All the combinations of the six features are evaluated to automatically classify the LP, VT, TR and OT events. Moreover, in contrast to Curilem et al. (2009) and San-Martin et al. (2010), the events are segmented in variable length windows to cover the entire event. Multiclass SVM classifiers are implemented by using one classifier for each class. Finally, a confidence step is proposed as an additional information to aid the classification process when the events are not well discriminated.

2. Materials and methods

The method adopted in this research can be divided into two main stages, preprocessing and classification, as shown in Fig. 2. This section describes each of the blocks within the classification system. The algorithm was developed in the Matlab environment, including signal handling, filtering, feature extraction, and pattern recognition. Furthermore, the built-in Matlab SVM functions were utilized to design the classifiers.

2.1. Signals database

OVDAS analysts chose the following three stations to provide the seismic signals for the paper: LAVE (−38.700988°; −71.651116°); LLAI (−38.784240°, −71.695261°); and PAILE (−38.873854°; −71.642720°). The different distances of the stations from the crater, as shown in Fig. 1, ensured adequate variability of the records. The

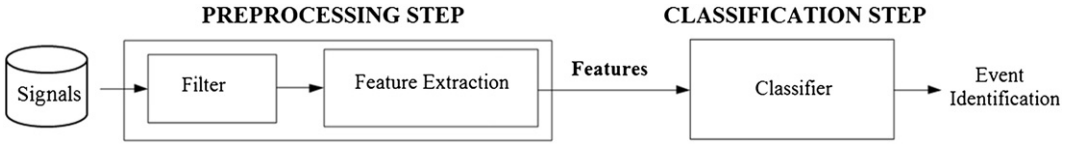

Fig. 2. General structure of the proposed pattern recognition system.

stations are broadband, Guralp 6TD of 30 s period. The signals were treated independently from their station. Moreover, only the Z component was considered because it provides a better signal/noise ratio in most of the events.

The events considered were LP, TR and VT and were recorded between 2009 and 2011. A fourth group, denoted as OT (other type), was defined to cluster all the signals that did not correspond to any of the first three events. OT includes background noise, tectonic events, rock falls, glacier defrost, and other events not related to internal volcanic processes. The purpose of creating the OT class was to increase the discrimination performance of the classifiers. This set is very important to discriminate signals that were not originated from inside the volcano.

The class and duration of each event were determined by an OVDAS expert. The database was generated by supervised selection of the events from the recorded data. Then, the selected signals were exported to the Matlab environment. A volcano seismologist also analyzed the data to confirm that the events were classified correctly, that is, data meet the classification criteria used by OVDAS to discriminate events from background. It is important to highlight that the length of the events stored in the database varies, as defined by the experts. Altogether, 1622 events of the four classes were stored in the database. The mean and standard deviation of the event durations, in seconds, are: LP, 55.2 ± 38.5; VT, 20.3 ± 9.3; TR, 147.5 ± 71.1; and OT, 93.4 ± 69.0. TR corresponds to longer events, thus the mean duration is higher. Duration of the OT events was less than 5 min with a relatively high standard deviation. An example for each group is presented in Fig. 3.

Time-frequency plots (spectrograms) of the considered four events of Llaima are illustrated in Fig. 4. Each spectrogram is obtained by normalizing the event with the maximum amplitude of its signal, and then divided into segments, each one being 10% of total length of the signal with 50% overlap. Each window is then transformed to frequency domain using the FFT function in Matlab to represent a slice in the spectrogram that covers the entire duration of a window. More details about the FFT can be found, for instance, in Oppenheim et al. (1999). It can be observed that all the events are of a different duration and spectral content.

The whole database was divided into three sets, i.e. training, validation and test. The training set contains the data used to adjust the parameters of the classifiers; the validation set is used to decide when to stop training, and the test set is used to determine the classification performance. Table 1 shows the structure of the sets.

2.2. Signal processing

Signals prior to 2011 were sampled at 50 Hz and after that at 100 Hz. Since the most relevant information for the volcano event classification

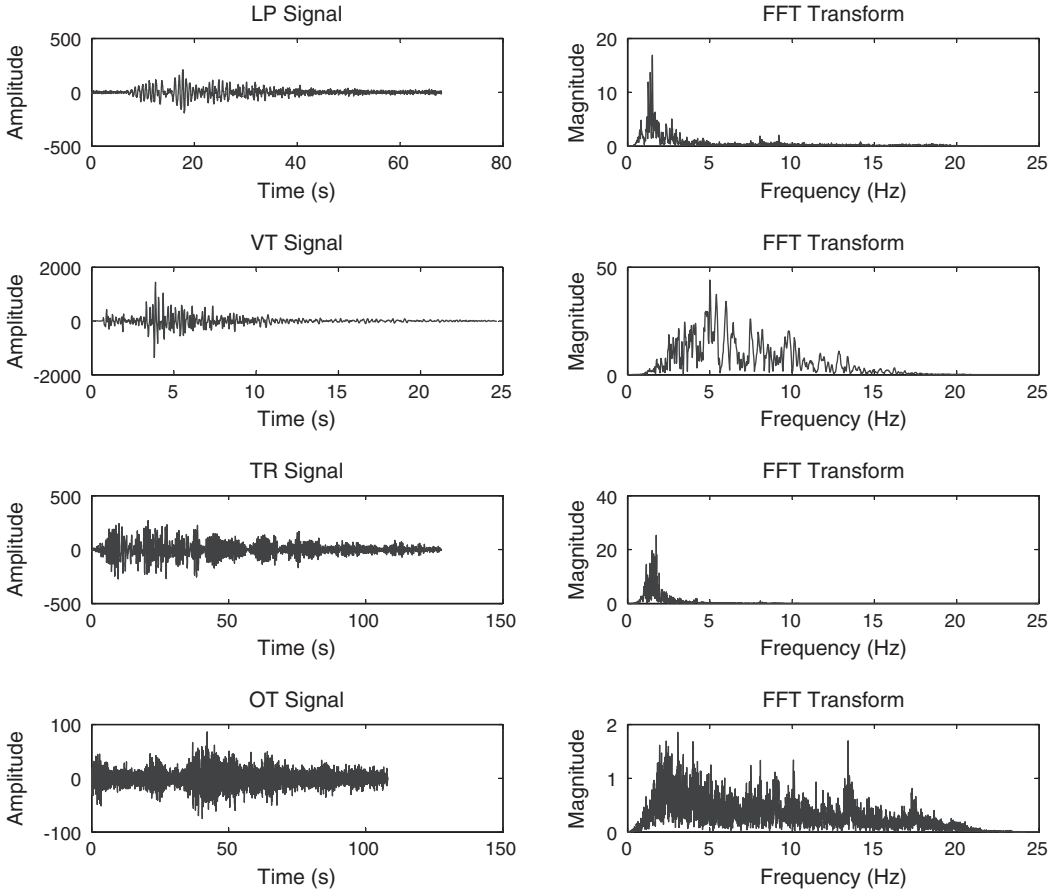

Fig. 3. Time and spectral representations of the signals belonging to the four classes of events being studied. From top to down: LP, VT, TR and OT.

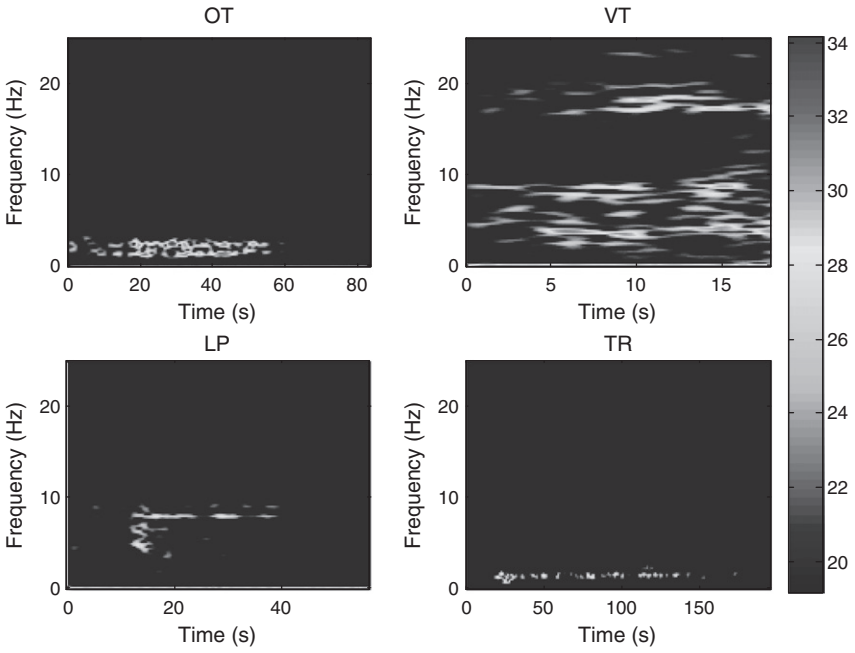

Fig. 4. Time-frequency representation of the four events: OT, TR, VT, and LP.

is below 25 Hz, signals sampled at 100 Hz were down-sampled to 50 Hz. All the signals were then filtered with an 8th order Butterworth band-pass filter between 0.5 Hz and 15 Hz.

Table 2 describes the features employed here: three of the considered features were extracted from the time domain representation of the signals; two parameters were obtained from the power spectrum; and, one feature was computed from the phase spectrum. The parameters were determined from variable length windows that covered the complete duration of events. All the extracted features were linearly normalized between −1 and 1.

The trigonometric first-order measure reflects the behavior of the phase part of the volcanic signal. The phase was obtained using the Hilbert transform in the range of [0 2π). Afterwards, the circular statistics was applied to obtain the statistical properties of this phase. As in the case of statistical linear dispersion measurements, sharpness symmetry underlying probability distribution can be defined from trigonometric moments. Several statistics were analyzed (i.e. mean, variance, skewness and kurtosis) but the moment of first order, given in Eq. (1), provided the best results:

$$\mu_1 = \frac{1}{K}\sum_{k=1}^{K} \exp(j\theta_k) \tag{1}$$

where K = 1,..., N is the number of samples and $\theta = \{\theta_k\}$ is the set of values in the circular range [0 2π) of the instantaneous phase of a random variable.

This complex number can be interpreted as the vector resulting from the sum of K unit vectors with angles given by θ. The resulting vector has magnitude $||\mu_1|| = 1$ and angle μ_1, called the θ direction, in the complex plane.

Table 1
Data distrubution.

Class	Training set	Validation set	Test set	Complete Data set
TR	176	163	161	500
VT	57	42	49	148
LP	156	182	161	499
OT	156	158	161	475
TOTAL	**545**	**545**	**532**	**1622**

Meanwhile, to obtain the relative energy per band the signal is decomposed by a wavelet transform. A Daubechies wavelet mother with a decomposition level equal to five was used. The percentage of energy was calculated using the following formula:

$$\text{Energy} = \frac{100 \times E_w}{E_{Total}} \tag{1}$$

where E_w is the sum of the components of the wavelet band, and E_{total} is the sum of all wavelet components (all the bands). The relative energy of all bands was tested, but the one calculated over the 4th band [1.56–3.125]Hz yielded the best results.

2.3. Classifier design

SVM is a classification approach which creates and defines the maximal margin hyperplane to separate two classes (Vapnik, 1995). An SVM algorithm endeavors to build a decision model and predicts whether a new point falls into one class or the other, as given by Eq. (3).

$$\mathrm{F}(x) = \sum_{i=1}^{\mathrm{N}} w_i . x + b = \sum_{i=1}^{\mathrm{N}} \alpha_i y_i (x_i . x) + b \tag{3}$$

where b is the distance between the separating hyperplane and the origin in the perpendicular direction, N is the number of support vectors, α is the non-negative Lagrange multiplier, y is the decision value $\in \{-1, 1\}$ and F is the decision function, which allocates a test sample to one class if its sign is positive, and to the other cloud if its sign is negative.

If the two classes are clearly separated, as shown in Fig. 5(a), the SVM model is designed as a linear division hyperplane with the greatest distance to the nearest points (i.e. support vectors) of each class (Burges, 1998; Hamel, 2009).

Since the events do not have in general linearly separable features, the linear hyperplane cannot be designed. The separation can be sought in an appropriately chosen kernel-induced feature space, as schematically shown in Fig. 5(b). The formula in Eq. (3) is modified as given in Eq. (4). Thus, the solution to the problem lies in finding the values of b, and α's.

Table 2
Overview of the extracted features used to define and classify each volcano's event.

Domain	Feature	Description
Frequency	E4	Percentage of energy in the 4th band [1.56–3.13] Hz, obtained from the wavelet transform of each event.
Time	Ax	Maximum value of amplitude obtained from the time records of each event.
Time	An	Average value of amplitude obtained from the time records of each event.
Time	Ad	Median value of amplitude obtained from the time records of each event.
Frequency	Fn	Average of the five highest frequency peaks obtained from the Fourier transform.
Frequency	Mc	Trigonometric first-order moment of circular statistics obtained from the instantaneous phase of each event.

$$F(x) = \sum_{i=1}^{N} \alpha_i y_i (\varnothing(x_i).\varnothing(x)) + b = \sum_{i=1}^{N} \alpha_i y_i K(x_i, x) + b \quad (4)$$

where $\varnothing(x)$ is the transform of sample x from non-linearly separable input space to a linearly separable feature space, and K is the applied kernel to represent this feature space.

Several kernel functions were investigated in constructing the required separating hyperplane between the two classes. The widely used kernel functions are the homogeneous polynomial of 1st, 2nd, 3rd, and 4th orders, and the Gaussian radial basis function (RBF) as given below. The RBF kernel function is presented in Eq. (5).

$$K(x_i, x) = e^{\frac{1}{\sigma}\|x_i - x\|^2}, \sigma > 0 \quad (5)$$

where σ is the Gaussian standard deviation.

An optimization process is used to find the values of (α_i, b) of Eq. (4), which correctly classifies samples with their associated class, maximizing the separating margin. The optimization process searches the solution given the parameter σ, which is the kernel parameter and c, a parameter that controls the trade-off between the complexity of the model and the accepted margin of error. The SVM functions of the bioinformatics toolbox of Matlab were used to solve the optimization problem. The RBF kernel was used as it presents many advantages compared to other common kernels, as exposed by Hsu et al. (2003).

Since SVMs are binary classifiers, they have to be combined to handle multiclass problems (Burges, 1998). A "one versus all" (also called "one versus rest") configuration is a simple combination in which each classifier discriminates between a positive class (e.g. LP), coded as 1, and all the other classes (e.g. VT, TR and OT together) are considered negative and coded as 0. The 0 is interpreted as the event not belonging to the actual class. So if N is the number of classes, N-1 is the minimum number of classifiers required to discriminate all the classes. However, in this work we propose a classifying structure formed by N = 4 classifiers, one for each class. Fig. 6 shows the resulting classifying structure.

In the "one versus all" structure the classifiers work independently, thus it may occur that many outputs are activated at the same time or none of them is activated. This occurs when more than one classifier considers the event as positive or when all the classifiers consider the event as negative (all the outputs are zero). These classification mistakes occur because all the classifiers are trained separately, thus when combined, the errors of each classifier contribute to the global structure error. Section 2.6 presents the method used in this paper to address this issue.

The output of the whole classification system was coded as shown in Table 3.

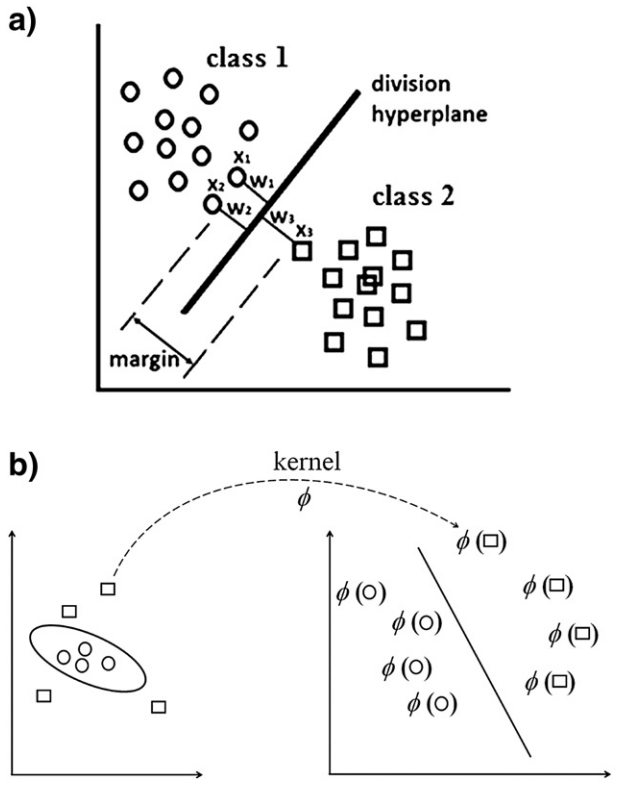

Fig. 5. (a): A classification problem of two classes (class 1, class 2), modeled by a linear SVM with three support vectors (x_1, x_2, and x_3) of three weights (w_1, w_2, and w_3), respectively. (b): A schematic display of how non-linearly separable samples are transformed from input space to feature space at which a linearly separating hyperplane can be designed.

2.4. Performance indices for the classifying structure

The Kappa coefficient is an index that measures the agreement of interobserver classification in multiclass problems (Landis and Koch, 1977), that is, when two or more independent observers are classifying the same things. The degree of agreement can be measured based on the number of classifications that was the same for all the observers. However, if they randomly assigned their ratings, sometimes agreement could just be due to chance. Kappa gives a numerical rating of the degree to which this occurs. The calculation is based on the difference between how much agreement is actually present (P_o, observed agreement) compared to how much agreement would be expected to be present by chance alone (P_e, expected agreement) (Viera and Garrett, 2005).

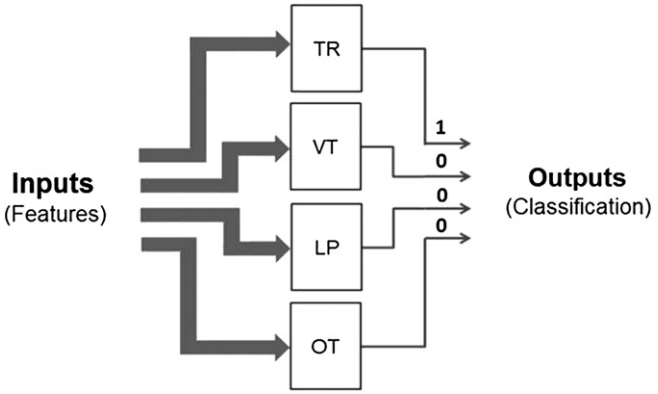

Fig. 6. Classifying structure: "one versus all" structure of the classifiers.

Table 3
Codification of the outputs.

Class	Classifier			
	1	2	3	4
TR	1	0	0	0
VT	0	1	0	0
LP	0	0	1	0
OT	0	0	0	1

In this work, as one judge is the expert and the other is the classifying structure shown in Fig. 6, the Kappa coefficient (K) evaluates the overall performance of the classifying structure. A value of K = 1 indicates complete agreement, which means that all the events were classified the same as by the expert. A value of K = 0 indicates no agreement, and negative values indicate that the coincidences are due to chance.

For C classes, the Kappa coefficient is computed by Eq. (6).

$$K = \frac{P_o - P_e}{1 - P_e} \tag{6}$$

where P_o is the observed agreement between the classifier and the expert, given in Eq. (7),and P_e is the probability that the agreement is due to chance and is given in Eq. (8). Further details can be found in Appendix A.

$$P_o = \sum_{i=1}^{C} p_{ii} \tag{7}$$

$$P_e = \sum_{i=1}^{C} p_{i\cdot} p_{\cdot i} \tag{8}$$

where p_{ii} is the joint proportion of the agreement and $p_{.i}$is the sum of the joint proportions of the classifier (rows) and the expert (columns); respectively, for each class. Between K = 0.41 and K = 0.6, there is a moderate agreement. Between K = 0.61 and K = 0.8 there is substantial agreement and from K = 0.81 to 1 the agreement is the highest (Viera and Garrett, 2005).

2.5. Individual classifier performance indices

To evaluate the performance of each classifier in the classifying structure, the contingency table was built for considering the "one versus all" structure, which considers the events of one class as positive and all the others as negative. Four statistical indices measure the performance of the individual classifiers: sensitivity (S_e), specificity (S_p), exactitude (E_x) and error (E_r), computed using Eqs. (9) to (12).

$$S_e = \frac{TP}{TP + FN} \tag{9}$$

$$S_p = \frac{TN}{TN + FP} \tag{10}$$

$$E_x = \frac{TP + TN}{n} \tag{11}$$

$$E_r = \frac{FP + FN}{n} \tag{12}$$

where TP (true positives) is the number of events correctly classified belonging to a specific class; TN (true negatives) is the number of events correctly classified as not belonging to a specific class; FP (false positives) and FN (false negatives) are the number of events classified erroneously. TP, TN, FP and FN were calculated from the contingency table. The statistical indices are determined for each model in all simulations.

2.6. Confidence measure

According to the scheme presented in Fig. 6, the final decision corresponds to the positive output of any of the classifiers. However, in the proposed system, the decision of each classifier is independent, and there are situations where the outputs of multiple classifiers are equal to 1, or all the classifiers generate zero. To tackle these scenarios, a confidence measure is estimated at the output of each classifier, as schematically described in Fig. 7. For each classifier output a reliability is assigned to assist in choosing the most reliable decision when there are more than one positive outputs or if all the classifiers output a zero.

As a confidence measure, the Bayes-based confidence measure (BBCM) is used (Yoma et al., 2005; Huenupan et al., 2008). This corresponds to the ordinary Bayes fusion weighted by the reliability of each individual classifier and provides a formal model for the reliability in a classification task.

The SVM classifier is a two-class classifier constructed with a sum of a kernel function, and the score is the distance between the sample and the hyperplane. If the sample is closer to the hyperplane, thus the decision of the classifier is less reliable. If $S_{CL_j}(X)$ is the score of classifier CL_j for an input feature X, with J the number of classifiers, in this case J is equal to four. The BBCM associated with classifier CL_j given a feature X is given by Eq. (13):

$$\mathrm{BBCM}\left(S_{CL_j}(X)\right) = \Pr\left(d_{CL_j} \text{ is ok} | S_{CL_j}(X)\right) \tag{13}$$

where "d_{CLj} is ok" denotes that the local decision of classifier j is correct. From Eq. (13) the probability of the classifier decision is correct (target or non-target) given the score $S_{CL_j}(X)$. By applying the Bayes theorem, Eq. (13) can be written as:

$$\mathrm{BBCM}\left(S_{CL_j}(X)\right) = \frac{\Pr\left(S_{CL_j}(X) | d_{CL_j} \text{ is ok}\right) \cdot \Pr\left(d_{CL_j} \text{ is ok}\right)}{\Pr\left(S_{CL_j}(X)\right)} \tag{14}$$

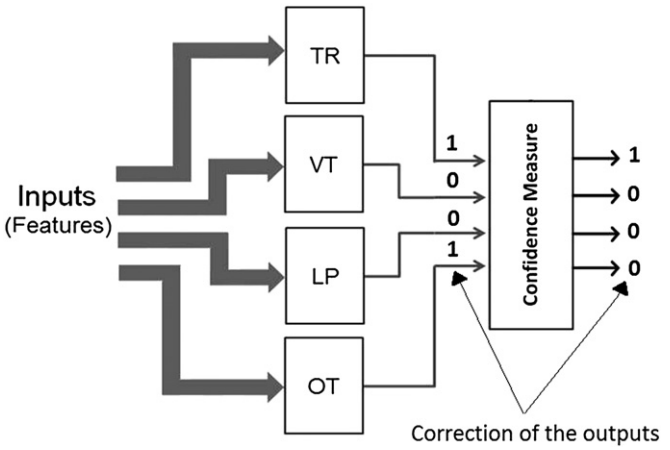

Fig. 7. Confidence measure scheme for the structure of classifiers. A correction of the outputs is performed according to the reliability of each classifier.

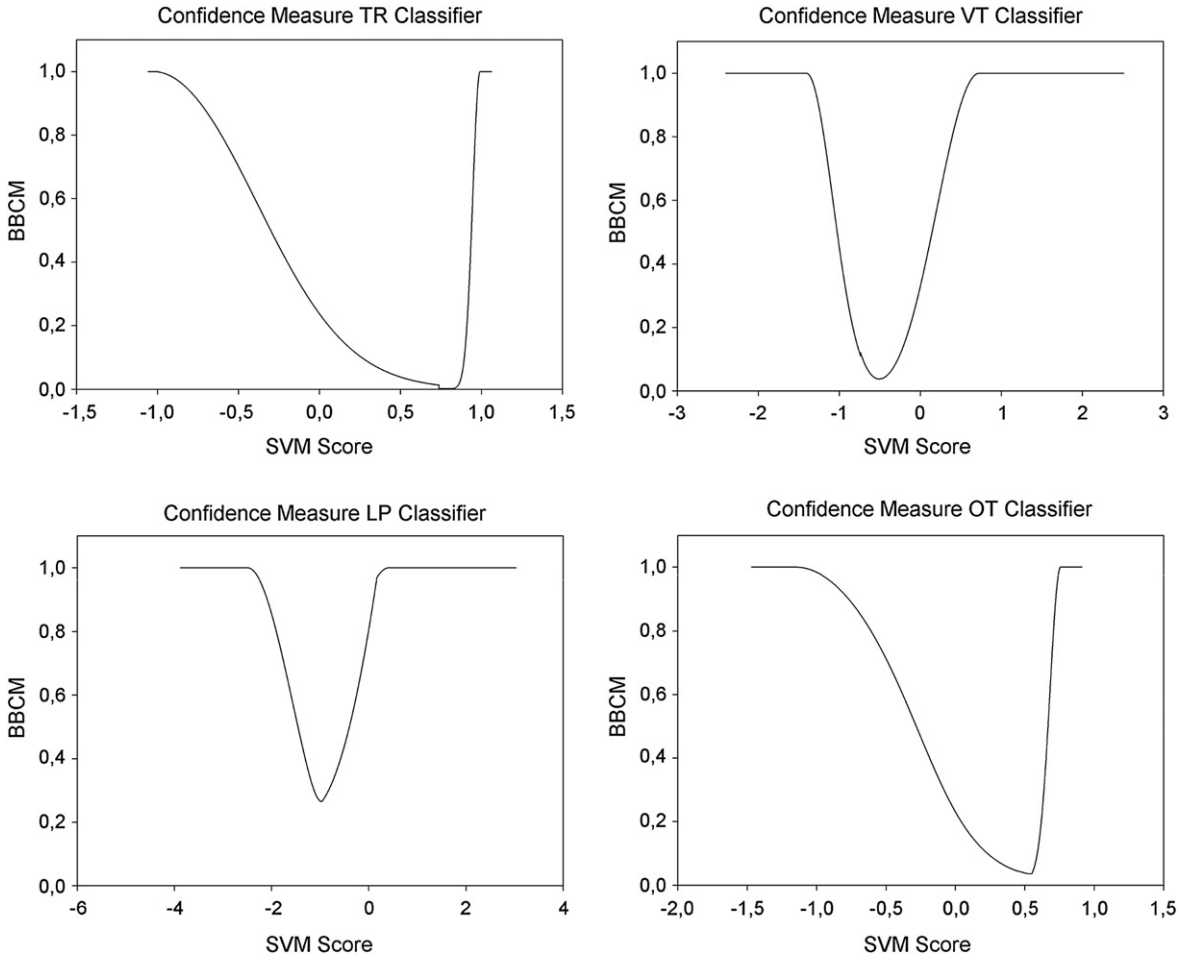

Fig. 8. Confidence measure curve with the best feature for each SVM classifier.

The BBCM$\left(S_{CL_j}(X)\right)$ is a probability itself. Moreover, the distribution $\Pr\left(S_{CL_j}(X)|d_{CL_j}\text{ is ok}\right)$ and the probability $\Pr\left(d_{CL_j}\text{ is ok}\right)$ provide information about the recognition engine's performance.

Fig. 8 shows the BBCM curves, as defined in Eq. (14), obtained for each classifier. A lower value of BBCM occurs when the classification score is close to the decision threshold.

3. Simulations and results

3.1. Features combination

A representation of the six features considered is provided in Fig. 9, where the values of each feature of the test set are plotted. On the horizontal axis the events are ordered by type and the vertical axis shows the amplitude of the normalized features.

Different combinations of these features define the input of the classifiers. The features were combined to evaluate which combination supplies the better classification performance. Fig. 10 illustrates the 63 possible combinations starting from individual features (combinations 1 to 6) to the last one that has all the features together (combination 63). It is important to highlight that each combination is used to train a classifying structure, where the four classifiers receive the same combination at their input.

3.2. Classifiers design

One classifying structure was trained for each of the 63 combinations of features, giving 63 classifying structures to evaluate and compare. This required the training of $4 \times 63 = 252$ classifiers. To implement each classifier, the design process of the SVM requires the tuning of the c and σ hyperparameters. The training set was used to adjust the hyperparameters of each SVM classifier. A grid search of 16×16 values was performed: the c and σ parameters were increased by powers of two: 2^x with $x \in [-5,10]$ and 2^y with $y \in [-7,8]$ respectively, with a step $= 1$ in both cases. Thus $4 \times 16 \times 16 = 1024$ SVMs were trained for the tuning process of each of the 63 classifying structures.

The validation set was used to compare the performances and to select the best c and σ combination. Here two main criteria were considered to select the classifier with the best performance: minimizing the validation error (criterion 1) and maximizing a fitness function given by the mean between sensitivity and specificity (criterion 2). Both results were analyzed. After the tuning processes, the ability of the best classifier to recognize new patterns was measured by evaluating the generalization performance with the test set.

The performance indices were obtained from the test set evaluation. We proposed to separately evaluate the performance of the whole classifying structure (joint evaluation) and the performance of each classifier (individual evaluation). The joint evaluation measures the agreement between the classifying structure and the expert. This evaluation uses the Kappa coefficient. In the latter each classifier was evaluated independently, according to its ability to identify the events of its positive class (sensitivity) and its ability to identify the events of its negative class (specificity). Thus the statistical indices were used in this case. The results are shown in the next two sections. When the best classifiers are selected for each event, the BBCM curves were estimated according to Eq. (14) with the same training data used to estimate the parameters for each expert.

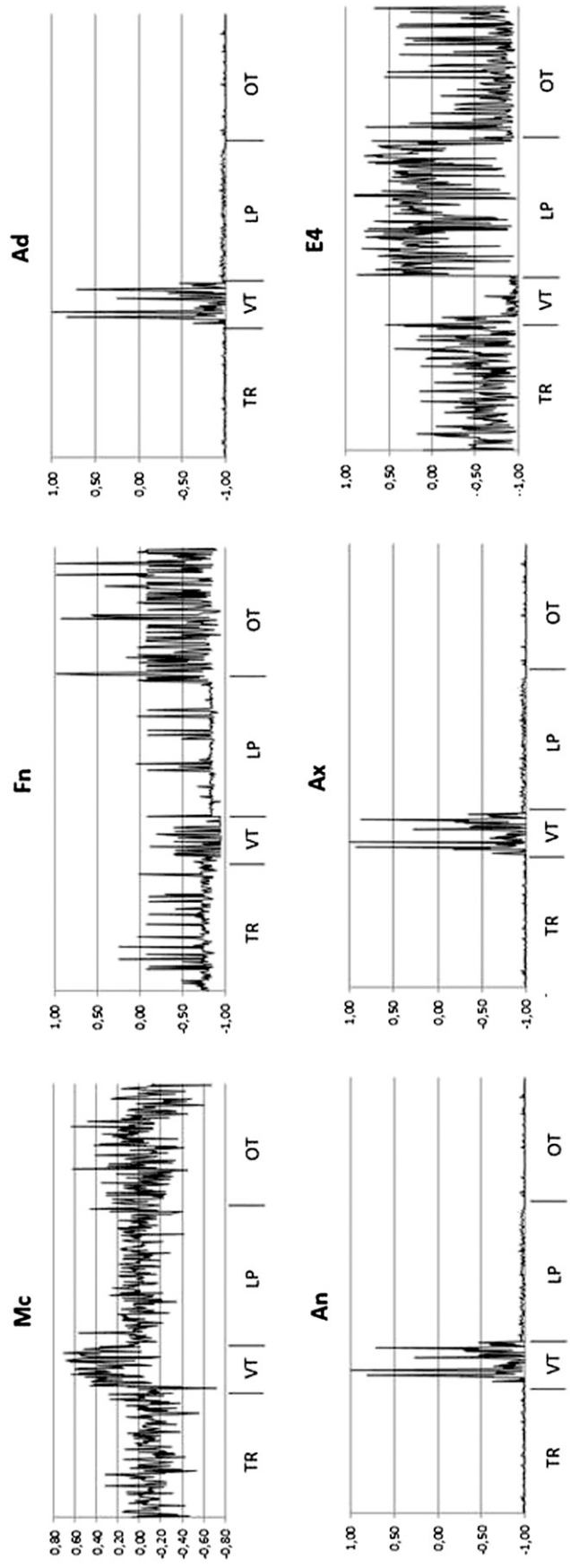

Fig. 9. Features values for the 532 events of the test set, for each class of event.

(1)	1						2														
(2)	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
E4						x											x	x	x	x	x
Ax					x								x	x	x	x					x
An				x						x	x	x				x				x	
Ad			x					x	x			x			x				x		
Fn		x					x		x		x			x				x			
Mc	x						x	x		x			x				x				

(1)	3																			
(2)	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41
E4											x	x	x	x	x	x	x	x	x	x
Ax					x	x	x	x	x	x							x	x	x	x
An		x	x	x				x	x	x				x	x	x				x
Ad	x		x	x		x	x			x		x	x			x			x	
Fn	x	x		x	x		x		x		x		x		x			x		
Mc	x	x	x		x	x		x			x	x		x			x			

(1)	4															5						6
(2)	42	43	44	45	46	47	48	49	50	51	52	53	54	55	56	57	58	59	60	61	62	63
E4						x	x	x	x	x	x	x	x	x	x		x	x	x	x	x	x
Ax		x	x	x	x					x	x	x	x	x	x	x		x	x	x	x	x
An	x		x	x	x		x	x	x				x	x	x	x	x		x	x	x	x
Ad	x	x		x	x	x		x	x		x	x			x	x	x	x		x	x	x
Fn	x	x	x		x	x	x		x	x		x		x		x	x	x	x		x	x
Mc	x	x	x	x		x	x	x		x	x		x			x	x	x	x	x		x

Fig. 10. The 63 combinations of the six features considered for classification. (1) indicates the number of features present in the combination, whereas (2) is the number of the combination.

3.3. Joint evaluation of the classifying structure

In this section the entire classifying structure is evaluated. Thus, the evaluation is performed comparing the joint agreement of the 63 classifying structures with the expert, calculating the Kappa coefficient for each structure. As explained in Section 2.3, a conflict occurs when more than one output is activated or when all the outputs are deactivated (0).

Three values of the Kappa coefficient were then considered to evaluate these situations:

1. Kappa min: is the worst value of agreement, because if there is no activation or if there are more than one activated outputs, the whole output is considered wrong in the contingency table.
2. Kappa max: is the best value of agreement because if one of the activated outputs is the right one, even if the other outputs are wrong, the output is considered right.
3. Kappa C: is the value of agreement when the outputs are corrected by the confidence block shown in Fig. 7. When there is no agreement, the confidence block activates the more reliable classification, correcting the output so only one classifier is activated.

Fig. 11 presents the Kappa min, Kappa max and Kappa C values of the different combinations, using different criteria to select the best classification. As shown in Fig. 11(a), combination 46 has the best Kappa value $C = 0.66 \pm 0.05$ ($\alpha = 0.05$) that means that Kappa $\in$ [0.61 0.71], while in Fig. 11(b), combination 46 is also the best combination with a Kappa $C = 0.65 \pm 0.05$, both considered good (Altman, 1991). The contingency table for this combination is presented in Table 4.

An analysis was made of the relationship between the number of features and the agreement. Table 5 and Fig. 12 show the Kappa C values according to the number of features used by the classifying structure. It is important to recall that in this analysis, the four classifiers of each classifying structure received the same combinations as input. Fig. 11(a) shows no significant increase of the agreement for more than two features (combination 9: Ad and Fn have a 0.659 ± 0.05 Kappa C, which is not significantly lower than the 0.662 ± 0.05 Kappa C value of combination 46). This is why one of the main results here is the importance of some features that reach high agreements alone, like Fn combined with another one like Ad or An. This implies that the frequency peaks and the amplitude median or mean are descriptive enough to reach a good agreement between the classifying structure and the expert. However, from the classification problem point of view, small but significant differences in classification agreement are valuable. The results show that only two features can be employed with a small degradation of classification accuracy. However, it is worth emphasizing that the increase in computational complexity required by adding new features is negligible, thus combination 46 remains the best combination.

Fig. 13 depicts the combinations that had the best Kappa C values (higher than 0.6). A detailed analysis of the combinations that yielded the best agreements shows again that the amplitude and the frequency peaks are the most relevant features. It can also be observed that some combinations of features decrease the agreement, like Fn and E4, both extracted from frequency. This combination reached 0.603, lower than the 0.618 reached by the Fn feature alone (combinations 2 and 18). This is also true when they are combined with other features like in combinations 46 and 62.

As can be seen in Fig. 13, the features considered by the classifiers with the best performance are mostly from amplitude (Ax, An and Ad) but the information of these features is redundant, as their combination gives no significant increase of the agreement. The first order trigonometric moment (Mc) is almost entirely absent in the best combinations. In addition to its poor agreement performance (0.1), combined with other features it mostly decreases the agreement, as

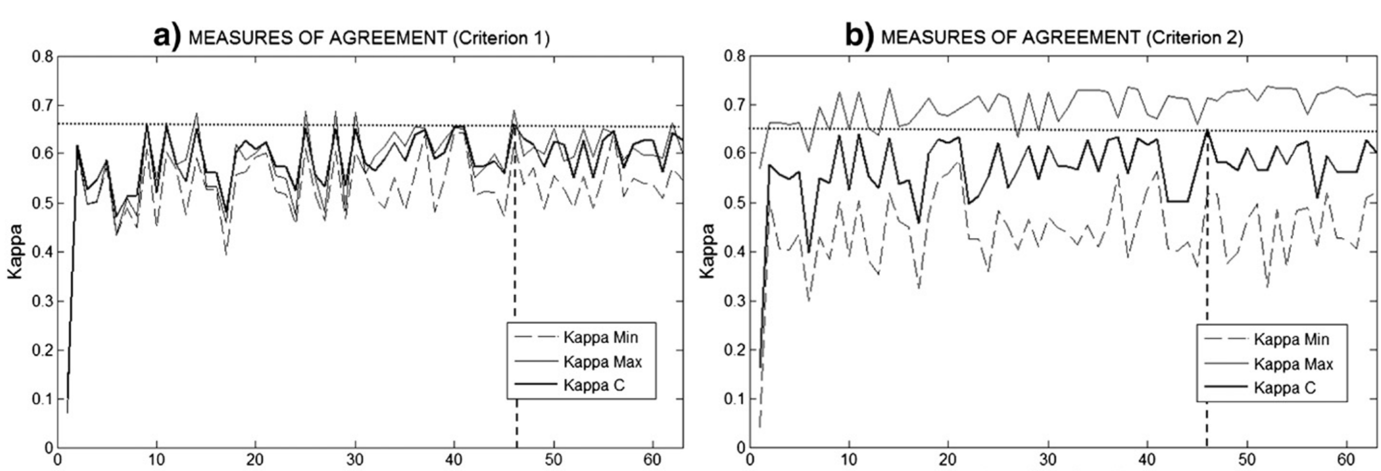

Fig. 11. Kappa values for the classifying structure of each combination. The classifiers were chosen using: (a) the validation error criteria (criterion 1), (b) the mean of sensitivity and specificity (criterion 2). The best combination (46) is highlighted in both figures.

Table 4
Contingency table of the classifying structure implemented with combination 46.

		Classifying Structure				
		OT	LP	VT	TR	Total
Expert	OT	111	17	3	30	161
	LP	16	139	2	4	161
	VT	5	0	42	2	49
	TR	26	21	3	111	161
	Total	158	177	50	147	532

shown when comparing the Kappa C values of columns 62 and 63 or 50 and 58 in Fig. 13.

3.4. Individual evaluation of each classifier

In this section the analysis is performed from the point of view of the class performance; thus, classifiers are analyzed individually. To identify the features that gave the best classifiers per class, the statistical indices were calculated, considering in each case the "one versus all" structure. So the contingency table is built considering all the events of a class as the true positives and all the events of the other classes as the true negatives. The statistical indices of the best classifiers and their input feature combinations are presented in Table 6.

TR events are identified using amplitude and frequency peak features. Their sensitivity is lower than the other classes, and the analysis of the contingency tables shows that the classifier tended to confuse TR with LP and mostly with OT. The energy of the4th band (1.56–3.13 Hz) is an important feature to discriminate VT from the other events. VT and OT have wide spectra, but different energy in the 4th band as noise has lower frequency components. The 4th band gathers the TR and LP events, and separates them from VT. Amplitude, as expected when analyzing Fig. 9, is also an important feature for discriminating VT. This class had the best performance with 98% of exactitude. LP had a good exactitude, 88% using only one feature extracted from the frequency peaks. OT events had a low exactitude and the analysis of the contingency tables shows that the classifying structure confused OT with LP and mostly with TR. The OT class is the only one that included the phase information as discriminator. OT is a very heterogeneous group of signals, thus it is expected that their discrimination will be more difficult, requiring features of a different nature such as the phase.

The best classifiers per class were chosen to implement a new classifying structure. In this structure, the classifiers received the different inputs shown in Table 6. Table 7 shows the new Kappa values for this eclectic classifying structure. It can be observed that it reached better Kappa C values, increasing the best level of agreement from 0.66 to 0.75.

4. Discussion and conclusions

The classic pattern recognition process supplies a clear sequence of steps to identify some events present in the seismic signals of a volcano. The main steps are preprocessing, which supplies the features extracted from the signals, and classification, which retrieves the class. The events are classified according to the information stored in their features, which is why features play a fundamental role in the whole process.

In this paper we presented a pattern recognition process to identify three important events of the Llaima volcano: the LP, VT and tremor, and to discriminate them from other events or noise present in the seismic records (OT). One focus of this study was to evaluate how the features that performed well in identifying events of the Villarrica volcano (Curilem et al., 2009) performed when applied to Llaima. Another focus of the study was to evaluate a circular statistics parameter extracted from the phase of the signals. This feature gave good results in a preliminary study (San-Martin et al., 2010) applied to Llaima. Thus, the present study analyzed six features extracted from the time amplitude of the signals, the frequency, and the phase.

It is important to underscore some differences between this work and previous ones. First, the classifiers were implemented using the SVM technique because the literature shows their good performance in seismic signal classification. Second, the events were cut in variable length windows. The experts at OVDAS indicated the beginning and the end of each event, and the whole window was considered for feature extraction. This is a fundamental difference because in the previous works fixed-sized sliding windows of 30 s (Curilem et al., 2009) or 60s (San-Martin et al., 2010) were considered to perform the feature extraction and then the classification. A variable size of the windows was considered because it is a more realistic situation for the automatic detection in the on-line application. Therefore, the classifying structure should receive the features extracted from the whole event. Finally, a confidence step was included in the general structure to solve classifying conflicts.

The classifier design considered two optimization criteria. This was done because the "one versus all" structure unbalances the classes: the positive class is almost one third of the negative three classes. This is critical when the positive class was the VT, because the number of positive events was 49 versus 483 negative events. These unbalanced sets affect the error measures; even when all the VT events were misclassified the error stayed very low. This is why the mean of sensibility and specificity was considered as a selection criterion for the classifier performance during the tuning process. The resulting classifiers were very different from the ones selected using the error criterion, as shown in Fig. 11. However, ultimately the best classifiers from both criteria had no significant performance differences (Kappa C value 0.75 versus 0.74).

An important outcome of this study is the effect of the confidence step at the output of the classifying structure. This step improves the performance of the classification as it defines the most reliable decision when the classifier structure is in conflict (more than one class or no class).

Analyzing the results, evaluations show good agreement between the classifying structures and the expert. The individual classifiers had more than 70% sensitivity and specificity, which is considered promising. The VT discrimination was the best, reaching 98% exactitude using the amplitude and the energy in the 4th band as discriminators. The LP discrimination was also interesting as it reached 88% exactitude

Table 5
Kappa C values for different numbers of features (rows) for the classifying structure of criterion 1.

1								0.10	0.62	0.53	0.55	0.59	0.47								
2				0.51	0.51	0.66	0.52	0.66	0.58	0.54	0.65	0.56	0.56	0.48	0.60	0.63	0.61	0.62			
3	0.57	0.57	0.52	0.65	0.56	0.53	0.65	0.54	0.65	0.58	0.56	0.59	0.62	0.59	0.64	0.65	0.59	0.60	0.65	0.65	
4				0.57	0.58	0.59	0.56	0.66	0.63	0.62	0.57	0.63	0.62	0.55	0.63	0.55	0.63	0.65			
5								0.57	0.62	0.63	0.63	0.56	0.64								
6										0.63											

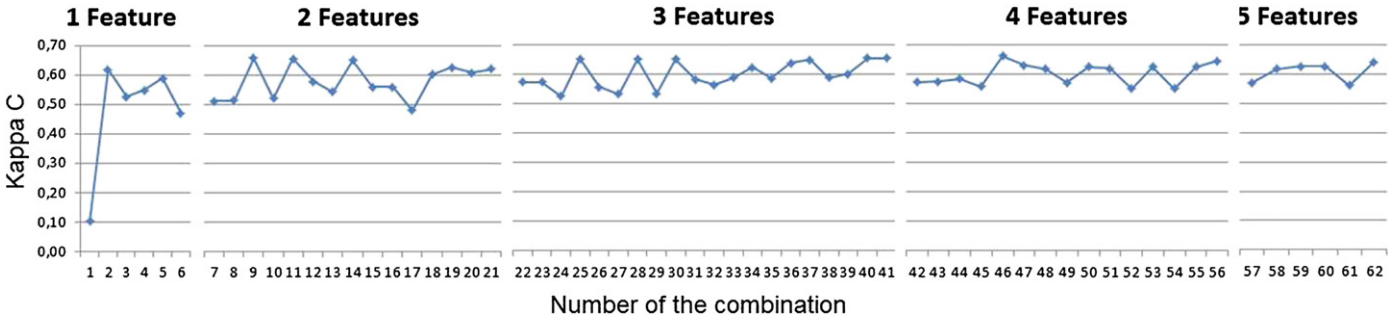

Fig. 12. Graphical representation of the Kappa C values for all the feature combinations.

with only one feature, extracted from the frequency peaks. The features considered in this work yielded very good results for these events. However, the most important discrimination problem remains the discrimination between TR–OT and TR–LP, because the classifying structure confused these three classes, mostly TR and OT. Llaima has been considered compositionally as a basic to intermediate volcano, with a reasonable prevalence of signals related to the fluid dynamic (mainly TR and LP events). This may explain the difficulty in separating LP and TR, due mainly to the similarity of their dominant frequencies. The difficulty indiscriminating OT is due to the kind of signals that comprise this heterogeneous group, composed mostly of noise that may have time amplitude and spectral energy similarities. For these signals we consider it necessary to continue the research looking for better discriminant features.

The circular statistics of the signal provided the poorest discrimination. It is important to note that the signals were filtered by a Butterworth filter which affects the phase within the bandpass. However the phase variation is linear with respect to the frequency, thus affecting all the signals in the same way: the phase patterns used in circular statistics are affected by the same amount, there by having no impact on the discrimination problem. San-Martin et al.(2010) found good results based on the phase information. However, in these works phase was extracted using a small number of samples, and perfectly aligned segments to compute the circular statistics. In this article, we considered a more realistic scenario, where segmentation and temporal analysis windows did not consider ideal segmentation conditions for the feature extraction. The results show that this has a strong impact on the phase characteristic, leading to a decreased performance by this feature. Furthermore, compared to the previous works, here an additional class was incorporated with a resulting further decrease in the effectiveness of the phase feature.

Analyzing misclassification of the whole system, we observe three issues that may explain it: first signals used to design and test the structures had noise as no signal/noise analysis was performed for creating the database. Second, the duration and the temporal behavior of the events were not considered. These features are important to discriminate TR and OT, the more misclassified classes. Third, the segmentation of the signals has to be improved as this work considered variable length segments, starting before and after the events, but without a common criterion to define its beginning and its end. All these issues will be tackled in future works. Treatment of noise may improve discrimination as classifiers will be trained with more accurate signals. A study on how different signal/noise ratios affect discrimination levels has to be performed. The duration and other temporal features of the events have to be studied as new features. For the on-line application of the system an automatic events' detection step has to be implemented to separate the events from the background signals, before the classification step. Then a uniform segmentation criterion has to be established. Finally, future works have to evaluate the information retrieved by different stations and other components of the signals and evaluate their impact in the performance of the classifiers.

An important advantage of the present work is that the identification system is performed with simple to compute features, which proved to be able to discriminate three important classes of volcanic events. This makes our paper propose a different approach compared to previous works. Amplitude, frequency and phase are features intrinsically involved in volcanic event analysis and their application to the classification of Villarrica and Llaima signals reached good

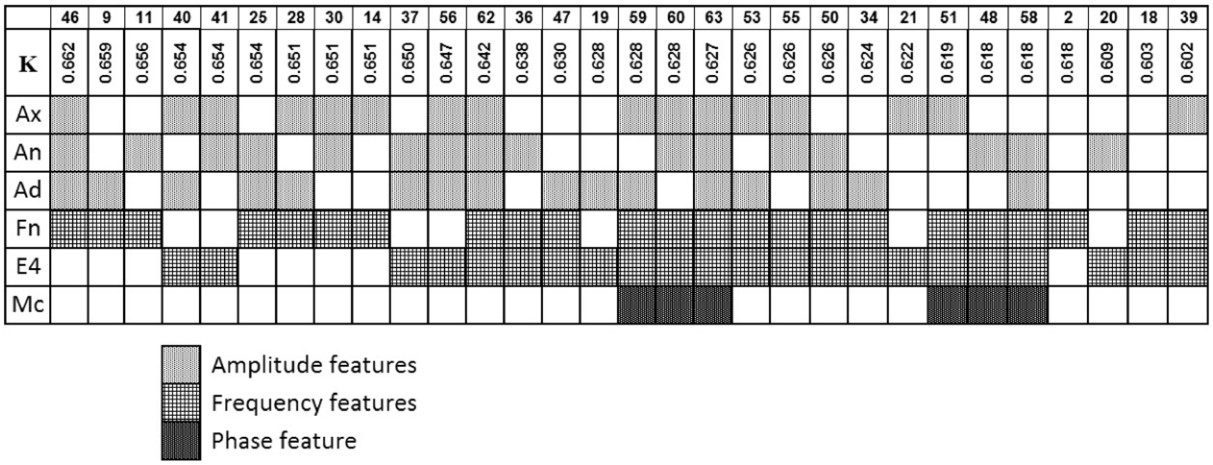

Fig. 13. Combination of features that retrieved the best Kappa C values (>0.6).

Table 6
Statistical indices and input features of the best individual classifiers of each class according to both classifier optimizations criteria. X means that the corresponding feature participates in the classification.

Optimization criteria	Generalization error				Mean of sensibility and specificity			
Class	TR	VT	LP	OT	TR	VT	LP	OT
Sensitivity	0.72	0.90	0.84	0.75	0.78	0.90	0.84	0.88
Specificity	0.89	0.99	0.90	0.84	0.79	0.99	0.90	0.71
Exactitude	0.84	0.98	0.88	0.81	0.79	0.98	0.88	0.76
Error	0.16	0.02	0.12	0.19	0.21	0.02	0.12	0.24
Combination	46	19	2	28	14	56	2	24
E4		X				X		
Ax	X			X	X	X		
An	X					X		X
Ad	X	X		X		X		X
Fn	X		X	X	X		X	
Mc								X

performances. Frequency is related to the location and shape of the seismic sources, which are different from one volcano to another. Although both volcanoes have similar structures and composition, the results are interesting because relevant frequency bands were the same, what was not expected a priori. Another advantage of the proposed method is that it does not model the signals as a function of the inner volcano structure. Consequently, one of the hypotheses assumed is to consider a volcano as a time invariant system. Signals used to train the classifiers have to reflect accurately the variability of seismic sources, the field of propagation and the characteristics of the recording stations, in one static context. The classifiers "learn" this variability and are able to generalize it to new signals, within the given context. However, seismic sources of each volcano are likely different and change according to its activity. If the inner volcano structure is modified (e.g. by an eruption), the pattern recognition models need to be re-trained, which in turn can be easily done. These advantages can make it possible to apply the method to dynamic volcanic situations and to other volcanoes.

The highest agreement reached was 0.75, which is a reasonable accuracy taking into consideration that the system had to discriminate between four classes, the features analyzed were very simple to compute and all the information of the windowed signal was processed without a detailed segmentation (realistic scenario). In conclusion, the paper shows that although further analysis is required to improve the results, as literature shows better performances, the simplicity of the proposed method and the future research lines defined by this work encourage us to improve this system. This is an interesting challenge for OVDAS since using signal processing in the detection of a volcano's events may provide a simple and cheap automatic recognition tool.

Acknowledgments

The authors would like to thank the project DIUFRO11-0032, the CONICYT-Anillo ACT1120, the FONDEF IDeA CA13I10273 and the FONDECYT 11110391 for financing the present work. Also many thanks to OVDAS, who provided the data and geological knowledge that supported the analysis of the results.

Table 7
Kappa C values for the classifying structure implemented with the best classifiers per class.

Optimization criteria	Generalization error	Mean of sensibility and specificity
Kappa min	0.71	0.58
Kappa max	0.78	0.84
Kappa C	0.75	0.74

Appendix A. Kappa coefficient

Inter-observer classification can be measured in any situation in which two or more independent observers are classifying the same things. The degree of agreement can be measured based on the number of classifications that were the same for all the observers. However if the observers randomly assigned their ratings, sometimes agreement should be just due to chance. Kappa gives a numerical rating of the degree to which this occurs (Viera and Garrett, 2005). The calculation is based on the difference between how much agreement is actually present ("observed" agreement, P_o) compared to how much agreement would be expected to be present by chance alone ("expected" agreement, P_e). This may be described in a contingency table, suchlike the one given in Table A.1 for C classes and two observers.

Table A.1
The contingency table for C classes and two observers.

		Judge 1			
		Class 1	Class 2	Class C	
Judge 2	Class 1	p_{11}	p_{12}	p_{1c}	$\sum \mathbf{p_{1.}}$
	Class 2	p_{21}	p_{22}	p_{2c}	$\sum \mathbf{p_{2.}}$
	Class C	p_{c1}	p_{c2}	p_{cc}	$\sum \mathbf{p_{c.}}$
		$\sum \mathbf{p_{.1}}$	$\sum \mathbf{p_{.2}}$	$\sum \mathbf{p_{.c}}$	**N**

For C classes, the Kappa coefficient (K) is computed by Eq. (A.1), where P_o is given in Eq. (A.2), and P_e is given in Eq. (A.3).

$$K = \frac{P_o - P_e}{1 - P_e} \tag{A.1}$$

$$P_o = \sum_{i=1}^{C} p_{ii} \tag{A.2}$$

$$P_e = \sum_{i=1}^{c} p_{i\cdot} p_{\cdot i} \tag{A.3}$$

Table A.2
Contingency table of combination 46.

		Classifying structure				
		OT	LP	VT	TR	Total
Expert	OT	111	17	3	30	161
	LP	16	139	2	4	161
	VT	5	0	42	2	49
	TR	26	21	3	111	161
	Total	158	177	50	147	532

Where p_{ii} is the joint proportion of the agreement and $p_{.i}$ is the sum of the joint proportions of the classifier (rows) and the expert (column), respectively, for each class. Example of calculation of the Kappa value is shown in Table A.2.

Observed agreement: $P_o = \frac{111+139+42+111}{532} = 0.758$

Expected agreement: $P_e = \frac{158*161+177*161+50*49+147*1611}{532^2} = 0.283$

Measure of the agreement: $K = \frac{0.758-0.283}{1-0.283} = 0.662$

Looking for this value in the qualitative evaluation table, i.e. Table A.3, it can be observed that this is a substantial agreement.

Table A.3
Qualitative evaluation table.

Kappa agreement	Evaluation
<0	Less than chance agreement
0.01–0.20	Slight agreement
0.21–0.40	Fair agreement
0.41–0.60	Moderate agreement
0.61–0.80	Substantial agreement
0.81–0.99	Almost perfect agreement

Appendix B. Confidence intervals calculation

The standard error of Kappa coefficient is given by:

$$SE(\mathrm{K}) = \sqrt{\frac{P_o(1-P_o)}{N(1-P_e)^2}} \tag{A.4}$$

For $\alpha = 0.05$, the $1-\alpha = 0.95$ confidence interval for K is given by:

$$I = 1.96 * SE(\mathrm{k}) \tag{A.5}$$

which means that K is defined in the interval $[\mathrm{K} - I\ \mathrm{K} + I]$ as it is considered normally distributed.

For the previous example $SE(\mathrm{K}) = 0.026$. Thus the interval is $I = 1.96 * 0.026 = 0.005$. The value of Kappa is then expressed as $\mathrm{K} = 0.662 \pm 0.005$. The confidence interval indicates that the population will have a K value inside the interval with a probability of 0.95 ($\alpha = 0.05$).

References

Altman, D.G., 1991. Practical statistics for medical research. Chapman & Hall/CRC Texts in Statistical ScienceChapman & Hall.

Álvarez, I., García, L., Cortés, G., Benítez, C., De la Torre, Á., 2012. Discriminative feature selection for automatic classification of volcano-seismic signals. IEEE Geosci. Remote Sens. Lett. 9 (2), 151–155.

Amari, S.-I., Cichocki, A., 1998. Adaptive blind signal processing-neural network approaches. Proc. IEEE 86 (10), 2026–2048.

Beyreuther, M., Carniel, R., Wassermann, J., 2008. Continuous hidden Markov models: application to automatic earthquake detection and classification at Las Canãdas caldera, Tenerife. J. Volcanol. Geotherm. Res. 176 (4), 513–518.

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

Cannata, A., Di Grazia, G., Aliotta, M., Cassisi, C., Montalto, P., P., Domenico, 2013. Monitoring seismo-volcanic and infrasonic signals at volcanoes: Mt. Etna case study. Pure Appl. Geophys. 170, 1751–1771.

Cannata, A., Montalto, P., Aliotta, M., Cassisi, C., Pulvirenti, A., Priviter, E., Patane, D., 2011. Clustering and classification of infrasonic events at Mount Etna using pattern recognition techniques. Geophys. J. Int. 185, 253–264.

Carniel, R., Barbui, L., Jolly, A.D., 2013a. Detecting dynamical regimes by self-organizing map (SOM) analysis: an example from the March 2006 phreatic eruption at Raoul Island. Boll. Geofis. Teor. Appl. 54 (1), 39–52.

Carniel, R., Jolly, A.D., Barbui, L., 2013b. Analysis of phreatic events at Ruapehu volcano, New Zealand using a new SOM approach. J. Volcanol. Geotherm. Res. 254, 69–79.

Curilem, G., Vergara, J., Fuentealba, G., Acuña, G., Chacón, M., 2009. Classification of seismic signals at Villarrica volcano (Chile) using neural networks and genetic algorithms. J. Volcanol. Geotherm. Res. 180, 1–8.

Chouet, B., 1992. A seismic model for the source of long-period events and harmonic tremor. Volcanic SeismologySpringer Berlin Heidelberg, New York, pp. 133–156.

Dowla, F.U., 1995. Neural networks in seismic discrimination. In: Husebye, E.S., Dainty, A.M. (Eds.), NATO ASI (Advanced Science Institutes) — Series E., 303. Kluwer, Dordrecht, pp. 777–789.

Erlebacher, G., Yuen, D.A., 2004. A wavelet toolkit for visualization and analysis of large data sets in earthquake research. Pure Appl. Geophys. 161, 2215–2229.

Esposito, A.M., Giudicepietro, F., D'Auria, L., Scarpetta, S., Martini, M., Coltelli, M., Marinaro, M., 2008. Unsupervised neural analysis of very long period events at Stromboli volcano using the self-organizing maps. Bull. Seismol. Soc. Am. 98, 2449–2459.

Falsaperla, S., Graziani, S., Nunnari, G., Spampinato, S., 1996. Automatic classification of volcanic earthquakes by using multi-layered neural networks. Nat. Hazards 13 (3), 205–228.

Gendron, P., Nandram, B., 2001. An empirical Bayes estimator of seismic events using wavelet packet bases. J. Agric. Biol. Environ. Stat. 6 (3), 379–402.

Giacco, F., Esposito, A.M., Scarpetta, S., Giudicepietro, F., Matinaro, M., 2009. Support vector machines and MLP for automatic classification of seismic signals at Stromboli volcano. Proceedings of the 2009 Conference on Neural Nets WIRN09, pp. 116–123.

Hamel, L., 2009. Knowledge Discovery With Support Vector Machines. John Wiley & Sons.

Hsu, C.-W., Chang, C.-C., Lin, C.-J., 2003. A Practical Guide to Support Vector Classification. Department of Computer Science, National Taiwan University.

Huenupan, F., Yoma, N.B., Molina, C., Garreton, C., 2008. Confidence based multiple classifier fusion in speaker verification. Pattern Recogn. Lett. 29 (7), 957–966.

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. Volcanol. Geotherm. Res. 187, 218–226.

Joevivek, V., Chandrasekar, N., Srinivas, Y., 2010. Improving seismic monitoring system for small to intermediate earthquake detection. Int. J. Comput. Sci. Secur. (IJCSS) 4 (3), 308–315.

Julian, B.R., 1994. Volcanic tremor: nonlinear excitation by fluid flow. J. Geophys. Res. 99 (B6), 11859–11877.

Lahr, J.C., Chouet, B.A., Stephens, C.D., Power, J.A., Page, R.A., 1994. Earthquake classification, location and error analysis in a volcanic environment: implications for the magmatic system of the 1989–1990 eruptions at Redoubt volcano, Alaska. J. Volcanol. Geotherm. Res. 62 (1–4), 137–151.

Landis, J.R., Koch, G.G., 1977. The measurement of observer agreement for categorical data. Biometrics 33, 159–174.

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, 1132–1144.

Langer, H., Falsaperla, S., Powell, T., Thompson, G., 2006. Automatic classification and a-posteriori analysis of seismic event identification at Soufriere Hills volcano, Montserrat. J. Volcanol. Geotherm. Res. 153, 1–10.

Lesage, P., Glangeaud, F., Mars, J., 2002. Applications of autoregressive models and time-frequency analysis to the study of volcanic tremor and long-period events. J. Volcanol. Geotherm. Res. 114, 391–417.

McNutt, S.R., 1992. Volcanic Tremor. Encyclopedia of Earth System Science. Academic Press, San Diego, California, pp. 417–425.

Messina, A., Langer, H., 2011. Pattern recognition of volcanic tremor data on Mt. Etna (Italy) with KKAnalysis — a software program for unsupervised classification. Comput. Geosci. 37, 953–961.

Mora-Stock, C., Thorwart, M., Wunderlich, T., Bredemeyer, S., Hansteen, T.H., Rabbel, W., 2012. Comparison of seismic activity for Llaima and Villarrica volcanoes prior to and after the Maule 2010 earthquake. Int. J. Earth Sci. 1–14.

Ohrnberger, M., 2001. Continuous Automatic Classification of Seismic Signals of Volcanic Origin at Mt. Merapi, Java, Indonesia, University of Potsdam, Postdam, Germany (168 pp.).

Oppenheim, A.V., Schafer, R.W., Buck, J.R., 1999. Discrete-time Signal Processing. Prentice Hall, Upper Saddle River, New Jersey.

Richardson, J.P., Waite, G.P., 2013. Waveform inversion of shallow repetitive long period events at Villarrica volcano, Chile. J. Geophys. Res. 118, 4922–4936.

San-Martin, C., Melgarejo, C., Gallegos, C., Soto, G., Curilem, M., Fuentealba, G., 2010. Feature extraction using circular statistics applied to volcano monitoring, progress in pattern recognition, image analysis, computer vision, and applications. LNCS 6419, 458–466.

Scarpetta, S., Giudicepietro, F., Ezin, E.C., Petrosino, S., Del Pezzo, E., Martín, M., Marinaro, M., 2005. Automatic classification of seismic signals at Mt Vesuvius volcano, Italy, using neural networks. Bull. Seismol. Soc. Am. 95 (1), 185–196.

Vapnik, V., 1995. The Nature of Statistical Learning Theory. Springer Verlag.

Viera, A.J., Garrett, J.M., 2005. Understanding inter observer agreement: the kappa statistic. Fam. Med. 37 (5), 360–363.

Vila, J., Macia, R., Kumar, D., Ortiz, R., Moreno, H., Correig, A.M., 2006. Analysis of the unrest of active volcanoes using variations of the base level noise seismic spectrum. J. Volcanol. Geotherm. Res. 153, 11–20.

Yoma, N.B., Carrasco, J., Molina, C., 2005. Bayes-based confidence measure in speech recognition. IEEE Signal Proc. Lett. 12 (11), 745–748.

Zobin, V.M., 2012. Seismic monitoring of volcanic activity and forecasting of volcanic eruptions, introduction to volcanic seismology. Elsevier, pp. 407–431.