ENHANCED CTBT MONITORING THROUGH MODELING, PROCESSING AND EXTRACTION OF SECONDARY PHASE INFORMATION AT HIGH SIGNAL FREQUENCIES

Eystein S. Husebye,[1] Yuri V. Fedorenko,[1] and Elena B. Beketova[2]

University of Bergen, Norway;[1] Institute of Geology, KRC RAS, Petrozavodsk, Russia[2]

Sponsored by Defense Threat Reduction Agency

Contract No. DSWA 01-98-C-0159

ABSTRACT

The outstanding problem in seismogram analysis is that of extracting more information on secondary phases at high signal frequencies. We need pP- and sP-phases for improved focal depth estimates while Sn and Lg can significantly improve epicenter determinations. But why are these phases so difficult to pick even by experienced analysts? Also, theoretical considerations imply that, for example, pP and sP should be far easier to observe than obviously is the case. One reason may be scattering in the lithosphere causing relatively strong P-coda that tend to mask secondary arrivals. However, the major problem is not only phase picking *per se* but also validation as genuine pP, sP and Sn-phase arrivals. Our preference here is for a probabilistic approach that is unsupervised Bayesian classification (preferable) or maximum likelihood solution for probability density functions that provide with basis in time tables (velocity model) and hypocenter (likely focal sphere orientation) the probabilities of observing these phases across a network of stations -a truly complex task. The first step in any event analysis is to ensure a reasonable epicenter location estimate, which are not always available from an international data center. We have developed simple, efficient methods for doing so given a single 3-component station or a small network of stations. In the latter case we a fit plane to P-arrival times and in this manner obtain bearing (azimuth), which, in combination with distance estimated from $t_S - t_P$ time, provides an accurate epicenter location. The next step is to record preprocessing in order to enhance pickings of secondary phases like pP, sP etc. and our set-up includes bandpass (Bessel type), wavelet denoising and polarization filtering. The Lg-phase is prominent at local distances but its onset is not clear so analysts often fail in feature extractions of these phases. Introducing the Hilbert transform in combination with Savitzky - Golay filtering, the corresponding envelope is smooth and exhibits a relatively consistent waveform. The aim of any processing scheme is to extract useful phase information from event recordings and here a major problem is phase validation. Except for Lg, there is a unique solution to this problem. Hence we tested the deformable template (DT) concept on local SeisSchool records from North Sea earthquakes. The goal is extraction of consistent Lg onset times and amplitudes for many combinations of events-stations and with excellent results. We have not explored the DT-potential for discrimination. In contrast to Lg waves, validation is a most severe problem in picking pP-, sP- and Sn-phases since there are no clear diagnostics for identifying these arrivals in the P-coda. To check for pP- and sP-phases we are using data mining techniques for converting seismograms into features like ragged peaks. Each peak is then a potential pP-, and/or sP-phase candidate or a noise wavelet. These peaks are subjects for simplified maximum likelihood classification as our first tool for phase validation. Ground truth (GT) information is a must in this context, but so far we use synthetics for simulating real seismograms at local distances including focal sphere sweeps of fault plane orientations that influence amplitude ratios like pP/P. Results so far imply that for a network of three stations there is a 50% probability of finding a focal depth diagnostic phase like pP or sP.

OBJECTIVE

A long-standing problem in observational seismology and in a nuclear-explosion or treaty monitoring context is that of accurately locating small events at local and regional distances. In particular, the focal depth parameter has proved troublesome because those phases most sensitive to depth, namely pP, sP, Sn and possibly Rg, are observationally elusive. The latter statement is somewhat ambiguous since there are many secondary phase candidates between the first arriving Pn or Pg and the dominant Lg or Sg. With the modern Comprehensive Nuclear-Test-Ban Treaty International Monitoring System (CTBT/IMS) seismic recording systems, phase picking *per se* is not much of a problem but the challenge is that of phase validation. Simply, how do we know that a secondary arrival in fact is pP? An analyst is severely handicapped here because the numbers of phase and candidate combinations for network recordings are truly large. There is no obvious criterion here for phase identification as the k-space location of pP is similar to that of the first arriving P-phase, and even expert analyst phase definition is seldom very helpful. However, secondary phases would be helpful in obtaining refined hypocenter estimates so our research objective is therefore phase validation of secondary phase arrivals in local recordings. Since traditional approaches to this apparently simple problem -- like analyst trace inspections, cepstrum analysis and fault space screening -- are not providing consistent and robust results, our preference here is for a probabilistic approach. A more comprehensive strategy is that of unsupervised Bayesian classification, but for now we test the concept using the simpler maximum likelihood solution for a probability density function that provides a basis in time tables and hypocenter (focal sphere) with the probabilities of observing these phases across a network of stations. The realization of this objective entails completion of several subtasks namely 1) access to good database(s) including some reliable ground truth events, 2) flexible signal processing schemes (beyond that of simple bandpass filtering and stacking of array records), 3) automatic extraction of secondary phase parameters and 4) natural expectations of time domain arrival intervals and strengths of pP, sP, Sn etc.

RESEARCH ACCOMPLISHED

Rome was not build in one day nor is validation of secondary phases accomplished in a year. The status of our research efforts here and associated accomplishments are listed below:

Databases for Research

With all the seismometers deployed at present, establishing proper databases should apparently not be much of a problem. Firstly, secondary phases are a mix of deterministic and random origins so an obvious advantage in the initial research stage is the ability to use the records from station clusters that are small networks. Station separations in the CTBT/IMS network are large so not so convenient when comparison and consistency of secondary phases are the issue. Anyway, we have asked several colleagues for access to their databases preferably including GT information for a few earthquakes but so far with little success. The only positive response we have received is from SAIC (Murphy pers. com.) from whom we have received some 10 Hindu Kush events. A minor problem here is that the crustal structure in this area is not accurately known. In other cases, the data set(s) available do not contain waveforms and even S-phase observations are lacking. We are continuing our efforts to find suitable waveform data bases for validation research.

SeisSchool Digital Seismograph Network

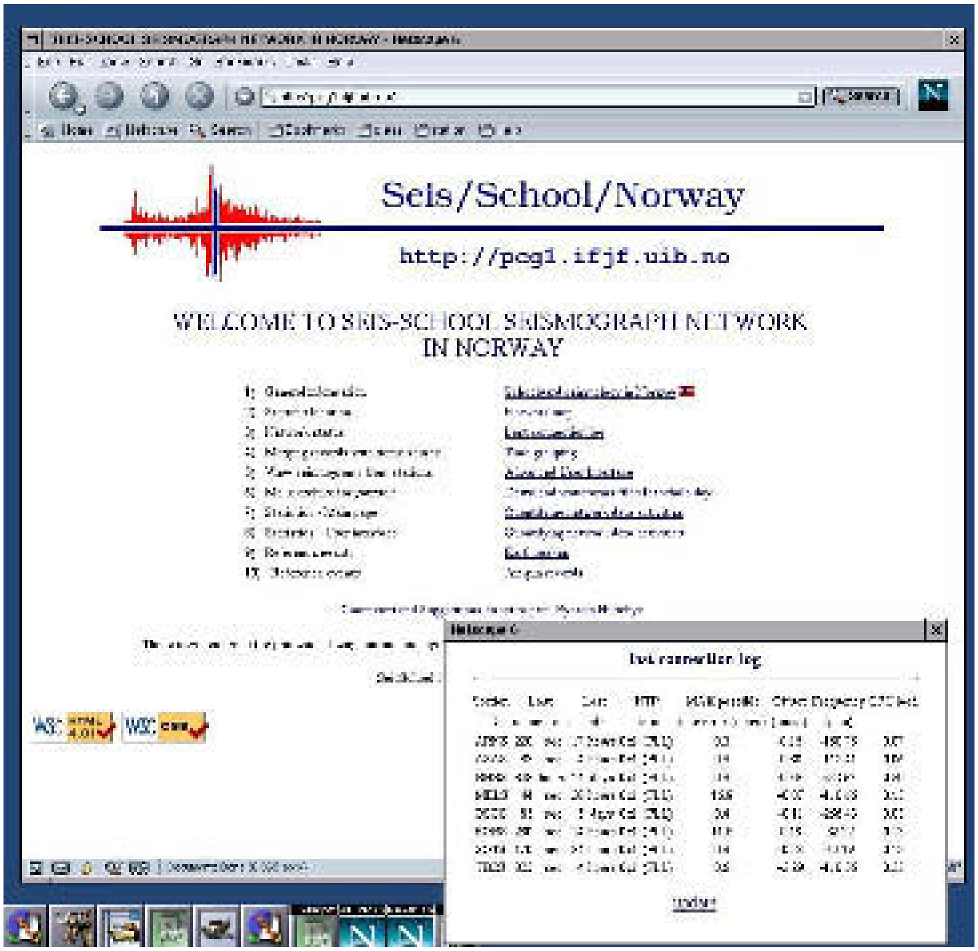
The initiation of these instrumentation efforts commenced roughly two years ago and the status now is that of seven operational stations (see our web page http://pcg1.ifjf.uib.no for details). There are no operational costs as data transfer is via Internet but of course the SP seismometers used, the Cossack Ranger, cost some $ 2000 each. Most stations are deployed in the Bergen area and local recordings provide data of uniform and high quality (Figure 1) that is most important for basic seismological research.

A prominent feature of our SeisSchool network is an advanced database structure allowing data storage and organizing in a manner similar to those at NMR and CTBTO and with advanced software schemes integrated in the database structure. Data analysis for school network recordings does not require special copyrighted software nor scientific/seismological backgrounds aiming at broad user segments -- try for yourself here.

The first step in any seismic data-related research is to obtain location and magnitude parameters for the event in question. For local networks, use of these data is an advantage only in case such information is not available from regional centers, NEIC etc. We have considered several approaches and our preference is for the one where

Table 1. Location experiment for North Sea earthquake 01/06/02/ at 06/18/10.0.[1]

Agency	Longitude	Latitude	Focal depth (km)
NORSAR	3.057 E	61.7902 N	15.9
NSN (Bergen)	3.124 E	61.7350 N	4.0
Our Network (manual)	3.433 E	61.9034 N	3.72
Our Network (envelop)	3.441 E	61.9028 N	3.61
Our Network (piecewise)	3.469 E	61.9445 N	3.70



Seis/School/Norway

http://pcg1.ifjf.uib.no

WELCOME TO SEIS-SCHOOL SEISMOGRAPH NETWORK IN NORWAY

Figure 1. Web-page presentation of our Seis/School/Norway network with insert showing operational status of network stations. Just click on 5) for inspection of network recordings including several filtering and trace zooming options. The first version of epicenter location for any network event can be tested on 9) Reference events -> Earthquakes.

[1] Our solution is based on four school stations of aperture ca 30 km and is close to those from NORSAR and NSN (Bergen) having many observational phases. Our phase picking was as described in the text; using original records and also feature extractions from envelopes. The focal depth estimate was critically dependent on picked Sn-phases.

also model parameters like Pg- and Pn-velocities are subject to estimation. The reason is simple; for 4 - 6 stations each providing 2-3 phase picks, we have enough observations for estimating more than the three standard hypocenter parameters. An example of locating a local earthquake is given in Table 1 with epicenter solutions from NORSAR and Bergen (national network) given as references.

Deformable Template Analysis - Envelope Waveform Matching

There are various 'joint epicenter location' schemes for improved source locations. A variant hereof would be to use deformable template techniques for matching envelope waveforms between different events in a given area or comparing waveforms between network stations. The idea is simple albeit computational complex, namely that for the jth event the recording at station X is a linear deformation (stretching) of that at station Y. Alternatively, we may have a cluster of events or aftershocks but with recordings from only one station but still undertake deformable template analysis with good results. Preliminary testing of this concept on records from our school network produced close similarity between reference and deformed station recordings as illustrated in Figure 2 by Husebye and Fedorenko (2001). Here we discuss its practical usage in terms of signal feature extraction. Firstly, the linear shift between the envelope traces X and Y can be converted to a relative distance estimate since the Lg-group velocity is well known. Alternatively, we use the reference station as a master station and read arrival time of Lg-max amplitude at a corresponding position on the other station traces. It is also feasible to compare in this manner records from different events and then Master Event location techniques may be introduced for very accurate epicenter determination.

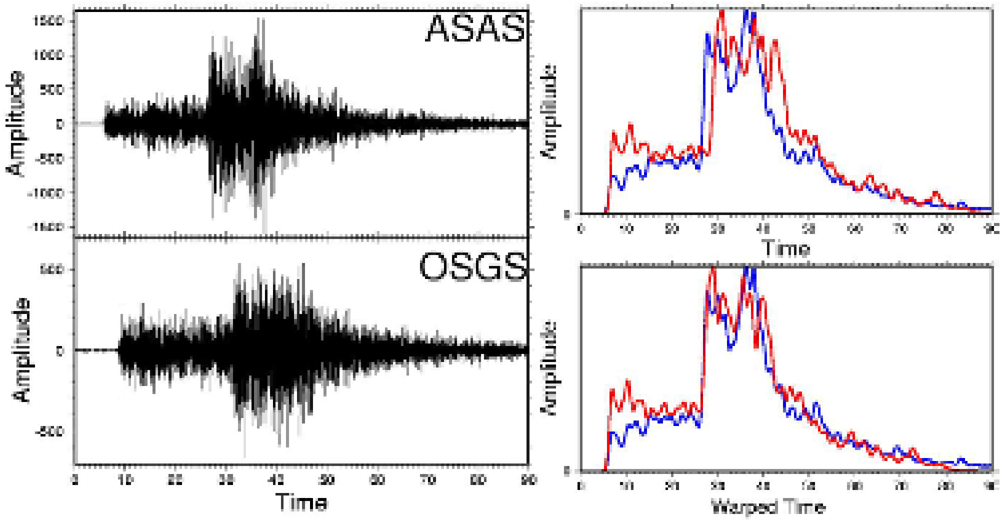


Figure 2. Linear time warping applied to envelopes of two Seis/School/Norway stations recordings.

An interesting research problem here is whether the deformable templates technique may be used for matching P-P-coda segments of records and thus improving secondary phase pickings and their subsequent validations.

Preprocessing Schemes

We pay much attention to the preprocessing scheme, because it produces phase candidates and implicitly defines resulting accuracy of the hypocenter location. We hope that automatic classification allows us to apply different signal processing procedures more freely because the probabilistic formulation of phase picking and hypocenter location would lead to automatic suppression of erroneous outliers. Filtering of initial waveforms in time domain is generally used to suppress that part of the spectra where the signal-to-noise ratio (SNR) is small, while in envelope processing, it is applied to smooth Hilbert transform output to suppress unimportant high-frequency noise and give extra weight to seismic phase arrivals. The same problem arises in polarization analysis where the phase is expected to be polarized but short in time; therefore the choice of a proper type of smoothing filter is essential. There is no general rule how to choose the best filter type and the optimal sequence of operations for an arbitrary signal. We discuss briefly our ongoing filtering tools preparations.

The most important part in building a preprocessing scheme is to understand how the information is contained in seismic signals. Regarding the problem of picking secondary phases, we concentrate on time domain processing and will select the filters by their pulse response, which describes how the information represented in the time domain is modified by filtering. We will consider Infinite Impulse Response (IIR) causal filters because of a previous bad experience with non-causal filters, which produced false ripples occasionally erroneously taken as phase arrivals.

Bessel, Butterworth and Chebyshev IIR Filter Types

There are many types of response approximation; we restrict ourselves to three commonly used filters: Bessel, Butterworth and Chebyshev. Bessel and Chebyshev types are two extremes; first is optimal for time domain and produces small artifacts (ripples) but is inefficient in frequency domain. The Chebyshev filter is "ringing" long after phase arrival thus producing a sort of artificial coda that may tend to obscure weak secondary phases. For this reason it is unacceptable for phase picking. The Butterworth type is an intermediate type, but it often produces ripples that can be taken as phase arrival, so we do not recommend it. As a rule, we are using the Bessel-type pulse optimal filter in view of its property of "preserving" signal energy within the first arriving cycles. Although Bessel-type filters are highly suitable for envelope smoothing, there is one interesting time-domain method of smoothing often named the "Savitzky-Golay filter". It is based on least squares polynomial fitting across a moving window within the data. A higher order polynomial makes it possible to achieve a high level of smoothing without attenuating the extremes in the data, which is essential for reliable secondary phase picking.

Raw Traces and Pre-whitening

Pre-whitening has a dual purpose of removing the low-frequency part of ambient noise and producing an approximate white noise spectrum as desired for the subsequent wavelet transform. Based on many spectral studies, we use a simple linear noise spectrum representation that is inversely proportional to frequency ω . Ambient noise exhibits, at least in coastal areas, strong seasonal variations in levels but not in form for $f > 1$ Hz, so the shape provides a robust and versatile spectrum representation. Using this noise feature we can easily transform initial seismic traces into traces with "white" seismic noise simply by differentiation preferably made through the Fast Fourier Transform (FFT). Stations in the SeisSchool network have a flat acceleration response; hence, pre-whitening is not used for our data, but it may be necessary to apply it to conventional sensors.

Wavelet Transform and Denoising

Like the Fourier transform, the wavelet transform (WT) is a linear operation on a data vector whose length usually is an integer power of 2, transforming it into a numerically different vector of the same length. In the wavelet domain basis functions $\psi(\tau, s; t)$ are somewhat complicated and have fanciful names like 'mother functions' and 'wavelets'. WT can be thought of as $X(\tau, s) = \int x(t)\psi_{\tau,s}(t)dt$ where $\psi_{\tau,s}(t) = \left(1/\sqrt{s}\right)\psi[(t-\tau)/s]$. In the WT domain the axes are scale s or inverse frequency and τ , which is the time shift as the wavelet slides through the signal window. Physically this means that low scales, which is the high-frequency part of the signal, has good time resolution while high scales are low frequencies with good frequency resolution. Hence, most of the usefulness of wavelet operation rests on the fact that $X(\tau, s)$ can be severely truncated as shown in Press *et al*. (1992), thus removing weakly represented scattering and noise contributions. This operation is especially useful for denoising non-stationary seismic waveform data. The choice and corresponding length of the wavelet function may not be overly critical; so far we have some experience with wavelets of the types Daubechies 4, 12 and 20 as shown in Daubechies (1988) and Press *et al*. (1992). In practical seismogram analysis we expect that in the case of pure noise with approximately flat spectra after pre-whitening, the wavelet transform will also be flat so wavelet coefficients for different s and τ will possess similar amplitudes, which may be removed without substantial distortion of the original signals. The ongoing experiments with sequential filter operations are encouraging; we clearly obtain better results than conventional bandpass Butterworth filtering. The above 'filter package' would be incorporated in the SeisSchool database to ensure much practical experimenting.

Secondary Phase Picking, Parameterization and Feature Analysis

Above we have outlined a number of signal processing schemes currently being tested for relative suitability on our SeisSchool recordings. Naturally, we consider these schemes well suited for preparing initial seismic traces for detecting weak secondary phases like pP, sP and Sn. To do so in practice, we have to quantify what we mean with signal features and naturally how to extract them from which type of records. Initially, we start with manual pickings of the arrival times of presumed secondary phases and then adapt the proper automated signal feature extraction scheme. In practice we do both in parallel as we need a data set of secondary phases for testing on maximum likelihood concepts of phase validations (next section) while here we briefly outline automated picking schemes. Firstly, at local distances, high-frequency signals dominate the seismogram but the recordings are complex and not well suited for automatic and flexible phase pickings. In our processing schemes the final trace output is, with few exceptions, the envelope trace and not the original waveform trace. Also, there are no clear wave theory anchored criteria for separating wavelets of respectively random and deterministic origins during the phase picking process. Since we cannot invert high-frequency signal waveforms, our feature extraction is tied to kinematical signal parameters like arrival time and amplitude in envelope traces.

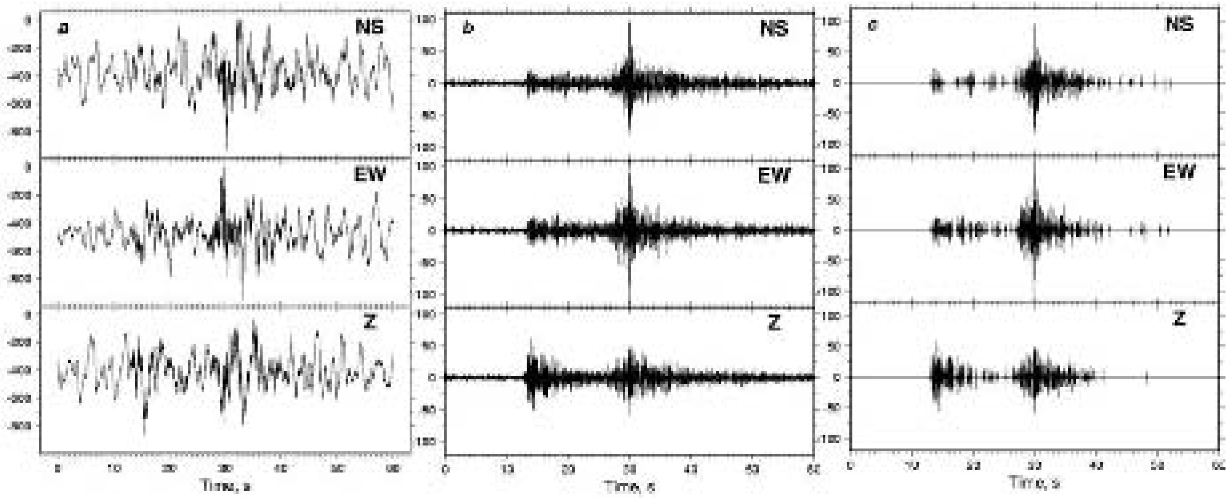


Figure 3. Original waveform processing. (a): Raw waveforms that are unfiltered Hoyanger (HYA) 3-component recordings of an arbitrarily chosen event; (b): Same event as in (a) but after pre-whitening; noise suppression is significant; (c): Same event as in (a) but after wavelet transform operation; preceding noise has been wiped out while the signal segments remains largely untouched.

Envelope Trace Segmentation

There are numerous techniques for economical time series representations as discussed by Pavlidis and Horowitz (1974) and Keogh and Smyth (1997) and methods for sequence recognizing and retrieval, where the features extracted are characteristic parts of the time series, so-called "feature shapes". Such methods do the same things as deformable template methods are supposed to do, that is time warping, shifting and uniform and non-uniform scaling of features. The latter is computationally much faster, partially due to economical representation of time series, partially because of simplicity of features. The economical representation critically impacts the sensitivity of the distance measure to various distortions and also can substantially determine the efficiency of the matching process. Thus, one seeks robust representation that are computationally efficient to work with. We are interested here in pursuing a representational language that can directly capture the notion of secondary phases shape. Piece-wise linear segmentations provide both an intuitive and practical method for representing curves in a simple parametric form (generalizations to low-order polynomial and spline representations are straightforward). There are a large number of different algorithms for segmenting a curve into the "best" piece-wise linear segments. Although appearing under different names and with slightly different implementation details, most of them can be grouped into one of the following three categories as shown by Keogh *et.al* (2001).

1. **Sliding Windows:** A segment is grown until it exceeds some error bound. The process repeats with the next data point not included in the newly approximated segment.
2. **Top-Down:** The time series is recursively partitioned until some stopping criterion is met.
3. **Bottom-Up:** Starting from the finest possible approximation, segments are merged until some stopping criterion is met.

Extracting of features from segmented envelopes is an easy and fast process. For example, we consider searching for the segments of the exact "inverted V" shape and/or rapidly ascending segments (Figure 4), using the angle between the adjacent segment as selector. We also extract the parameters of a segments constructing feature. The final choice of which shape "best" represents secondary phases is a major problem that must be resolved using simulated traces, real recordings from our network and GT recordings at local distances.

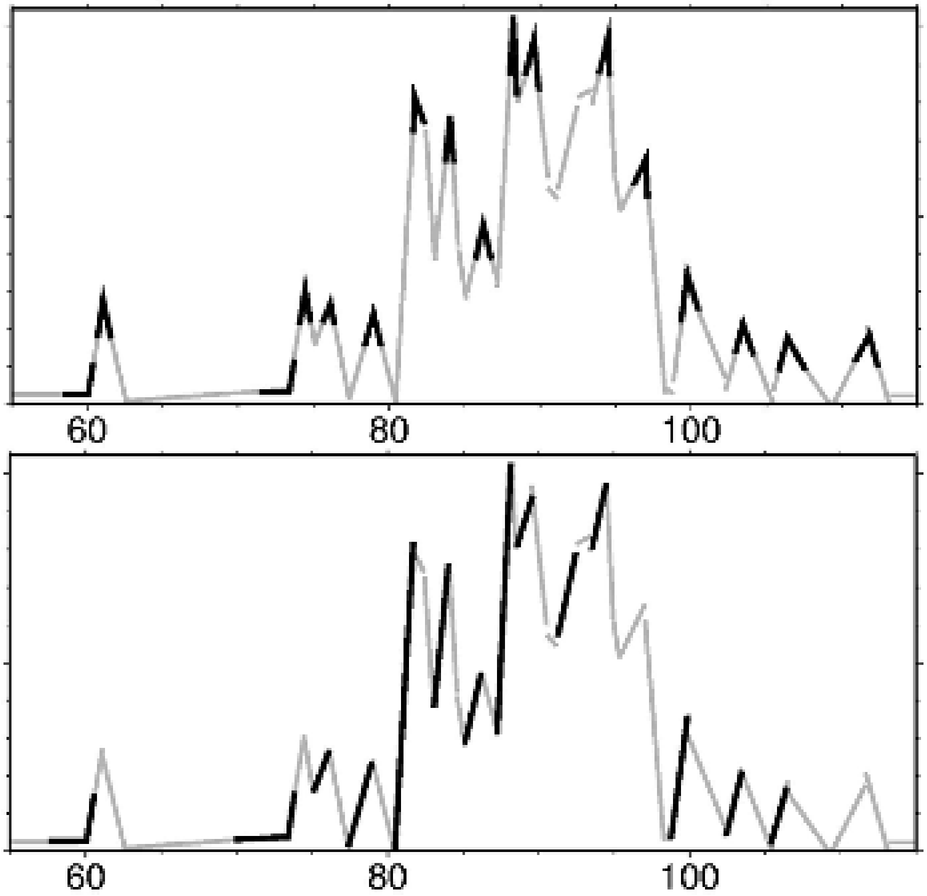


Figure 4. Extracting of features from envelope traces; upper triangle shapes and lower steep lines are used in our data mining experiments on trace segment representations. Both of these schemes give largely similar feature representations. Horizontal axis: time in seconds.

Needless to say, much experimenting is needed in determining which trace segmentation scheme is adequate for extracting phase information in an automatic manner from the seismic record.

Statistical Approach to Phase Validation

The problem of properly picking secondary phase arrivals is encountered in many seismological disciplines, perhaps foremost in long-range seismic profiling investigations. It appears that many of our colleagues by definition presume that any secondary arrival stems from a clearly defined crustal discontinuity stretching laterally for hundreds of kilometers -- both from a physical point of view, as this is not likely due to small impedance contrasts that can be accommodated in the crustal wave guide, but also from a geological point of view due to many tectonic deformations in the past. Also, seismic analysts are occasionally eager to report many (presumed) prominent phases, which regrettably do not comply with travel-time expectations from standard Earth models. Hence, for a start, we may assume that the wavelet signal amplitudes are large relative to ambient noise and the P-coda so major problems in phase picking and phase identification originate from random layer reflections and complex topography. In other words we have more phase candidates than the obvious three classical ones -- pP, sP and Sn -- arriving between Pn or Pg and Lg or Sg. The phase readings stem from a network of five to ten stations so we are faced with a formidable multi-dimensional data space of observations to analyze with respect to arrival times (velocity model) and source radiation pattern (amplitude ratios).

Intuitively, we may wonder if the size of this data space can be reduced by more optimal processing schemes and/or by a trained analyst to identify false phase wavelets. With all combinations at hand such an approach would be futile in our opinion. However, a more general mathematical approach is to formulate the above validation problem as a classification or 'clustering' problem. A deterministic, brute force solution would be to search the whole model space, thus assigning the detected phase candidates to separate classes that conform to expected arrival times and amplitude ratios for pP and sP. Again, we are not convinced that this strategy really would provide a good solution to our validation problem. Below, we outline our approach to the phase validation problem using a maximum likelihood formulation and classic statistical mixture model for clustering. Although ML clustering is less general than Bayesian unsupervised classification, we sacrifice generality for the sake of simplicity. Mixture modeling concerns modeling a statistical distribution by a weighted sum of other distributions.

Let the input vectors $\mathbf{x}_1$, $\mathbf{x}_2$, $\ldots$, $\mathbf{x}_n$ be observations from a set of unknown distributions E_1, E_2, $\ldots, E_k$. Suppose that the probability density of the observation $\mathbf{x}_r$ with respect to E_i is given by $f_i(\mathbf{x}_r \mid \theta)$ for some unknown set of parameters θ. Also, suppose for each $\mathbf{x}_r$ and distribution E_i τ_r^i represents the probability that $\mathbf{x}_r$ belongs to E_i. Each input is constrained to belong to some distribution, so $\sum_{i=1}^{k} \tau_r^i = 1$. Given these definitions, the goal of the scheme is to find the parameters θ and τ_r^i that maximize the likelihood

$$L(\theta, \tau) = \prod_{r=1}^{n} \sum_{i=1}^{k} \tau_r^i f_i(\mathbf{x}_r \mid \theta)$$

Following Banfield and Raftery (1993), we simplify this equation even more by assuming that each input must belong to exactly one distribution. It discards the probabilities τ_r^i and introduces the vector $\boldsymbol{\gamma} = (\gamma_1, \gamma_2, \ldots, \gamma_n)$ to represent the identifying label for observations, so $\gamma_r = i$ if $\mathbf{x}_r$ comes from E_i. The goal is then to find the parameters θ and γ that maximize the likelihood

$$L(\theta, \gamma) = \prod_{r=1}^{n} f_{\gamma r}(\mathbf{x}_r \mid \theta) \tag{1}$$

Now the problem is to define $f_{\gamma r}(\mathbf{x}_r \mid \theta)$ for each possible distribution E_i using our knowledge about velocity model, hypocenter and stations locations, fault plane orientation, lithosphere scattering properties and surface and Moho reflectors that define P-coda duration, variance of ambient seismic noise and seismic signal amplitude.
As a practical example; we consider the simple case of the p.d.f. in $f_r(\mathbf{x}_r \mid \theta)$ only relating to wavelet arrival time. We further assume that all possible hypocenters are located within an elliptically shaped cylinder with discrete hypocenter positions O_i. Let the number of stations be Q and these stations record all M seismic phases. For each O_i and station S_q we calculate arrival time for M phases $\mathbf{T}_q = T_{iq}^m$, $m = 1, \ldots, M$. The measured arrivals from O_i to station S_q are $\mathbf{t}_q = t_{iq}^m$ are presumed normally distributed with mean t_{iq}^m and variance σ_{iq}^m. Then, at each station we compose $\mathbf{t}_q$ choosing M phases from K candidates, $K \geq M$. Finally, we find γ which maximizes (2); it would be the set of secondary seismic phases expected from the model.

Naturally, it may be argued that this example is too simplified. Firstly, it is not likely that all phases appear between candidate wavelets at each station. Some phases obviously will be missed at some stations so we need to define the probability or compose our cases with missed arrivals. On the other hand, the unsupervised Bayesian learning approach is able to treat such cases and thus is superior to the simpler likelihood formulation on this account. Secondly, information about relative phase amplitudes, which really controls accuracy of measurements and defines how likely it is to miss a phase, was not included in the above model as illustrated below. Still, the above example could be instructive. For example, we desire to estimate the probability that the power ratio $R = P_{\text{depth}} / P_{\text{Pg}}$ is greater than, for example, 10 at focal sphere; P_{depth} is a power of a depth phase leaving the focal sphere along the ray connecting source and station. In this case we presume that the Pg coda will not mask depth phases and that the fault plane orientation has a uniform angular distribution. The three recording stations are 130 km away from the epicenter with azimuths 0, 30 and 60 degrees respectively. Table 2 summarizes results of evaluating secondary phase power relative Pg obtained by a Monte-Carlo test. *N* is the number of stations involved; "Any phase" column shows the probability that *R*>10 for at least one secondary phase at focal sphere for any station.

Table 2. Probabilities for finding secondary phases pP, sP and Pmp in the earthquake records assuming uniformly distributed fault plane orientation.

N	$P(R_{\text{Pn}} > 10)$	$P(R_{\text{pP}} > 10)$	$P(R_{\text{sP}} > 10)$	$P(R_{\text{PmP}} > 10)$	**Any phase**
1	0.14	0.13	0.10	0.14	0.37
2	0.28	0.27	0.20	0.27	0.50
3	0.43	0.39	0.30	0.41	0.61

Table 2 implies that by using three stations, we may expect to pick at least one of the secondary phases masked by a strong Pg-coda with a probability about 60%. These results are also incomplete because the positions of secondary phase arrivals relative to the strong Pg phase are also important.

CONCLUSIONS AND RECOMMENDATIONS

Seismology provides a unique, observational platform to the dynamic Earth by recording elastic waves generated by small and large earthquakes any place in the world. As a science, seismology is exclusive since instrumentation is very expensive and subsequent data accesses are problematic for non-seismologists as well as many other geoscientists. We have here demonstrated that design and construction of inexpensive but high-quality SP seismometers are feasible and that easy, cost-free data access via the internet is now feasible (Figure 1).
Our end product will be a comprehensive phase picking and phase validation scheme. This would imply that analyst participation in routine record analysis and bulletin preparation work will be significantly reduced. In our opinion the initial waveforms are too complex for automated phase picking so we start analysis with simplifying waveforms and enhancing phase arrivals. These preprocessing steps include waveform bandpass filtering, then constructing rectilinearity and planarity envelopes using eigenvalues of polarization matrix and finally making a piece-wise envelope approximation. In order to estimate how much these transformations affect the final location, we compare the hypocenter position obtained from a) phases picked manually from original waveforms; b) phases picked manually from envelope maximum and c) obtained from "inverted V" features (Figure 4) of piece-wise approximations.

The results are slightly different, but the maximum distance between various epicenter locations for an off-shore earthquake in western Norway was less than 8 km and focal depth differences were within 2 km. To us, these results are encouraging at this initial step of concept development and the applicability of our preprocessing scheme. Step by step we proceed in the following manner:

1. We automatically obtain a first approximate epicenter location from first arriving phases through 'wavefront fitting'. Using the shape "like _/" we obtain first arrivals from piece-wise envelopes for all stations. At this stage we need to test whether these first arrivals are Pg or Pn or a Pg and Pn mixture across the network. We relocate and compare the residuals; if residuals are large, we rename Pn <-> Pg until location converges. Then we add one clearly seen phase Sn, Sg or Lg type and obtain an improved epicenter location.
2. We set the uncertainty of phase arrival time, say 0.3-0.5 sec; surround the epicenter by a mesh and use a grid search method to obtain the maximal value of the likelihood function. Then we add one of phase candidates for, to say, Sn, and recalculate likelihood. The candidate that adds most to the likelihood function is chosen as this stage.

3. Recalculate the hypocenter and repeat step 2 until we have a stable solution for hypocenter. We are now in the process of analyzing many local events -- not all from our own network -- and a particular interesting aspect here is whether we could estimate realistic probabilities of observing and picking secondary onsets bearing on focal depths.

Future Research: Locating accurately small seismic events recorded at local and regional distances, and particularly improving focal depth estimates, remains an outstanding problem. Our own research strategy and hence recommendations are:

1. More reference data sets stemming from dense networks and some events must comply with Ground Truth requirement for GT5 at least.
2. More attention must be given to the way records are analyzed; secondary phases like pP, sP and Sn are not obviously identified after simple bandpass filtering. Polarization filters may suppress random wavelets. Such phase observations are important for improved focal depth estimates.
3. Since secondary phases are not easily picked, probabilistic phase validation schemes are important as demonstrated here. We need far more observations in order to explore the most advanced strategies.
4. Learning aspects: deformable wavelet matching. In this way source location distortions are easily demonstrated but whether the technique is useful for validating deterministic secondary phase pickings remains to be seen.

Final remark: improved feature extractions from local recordings and hence improved CTBT monitoring capabilities dictate non-conventional approaches in both recording analysis and probabilistic interpretational tools in view of modest research progresses in this field over the last decades.

REFERENCES

Daubechies, I., (1988). Orthonormal bases of compactly supported wavelets. *Communications on Pure and Applied Mathematics,* **41**, pp. 909-996.

Fedorenko, Yu. V., E. S. Husebye and B. O. Ruud, (1999). Explosion Site Recognition: Neural Net Discriminator Using Three-Component Stations, *Phys. Earth Planet. Int*., **113**, 131-142.

Husebye, E. S. and Y. V. Fedorenko, 2001. Seismic regionalization, signal detector and source locator. Proceedings of the 23rd DOD/DOE Seismic Research Review: Worldwide monitoring of Nuclear Explosions. Rep. no. LA-UR-01-4454. Vol. 1, 242 - 251. Oct 2-5/01, Jackson Hole; WY, USA.

Fedorenko, Yu. V. and E. S. Husebye, (1999). First Breaks - Automatic Phase Picking of P- and S- onsets in Seismic Records. *Geophys. Res. Lett*., **26**, 3249-3253.

Keogh, Chu, Hart, and Pazzani, (2001). An Online Algorithm for Segmenting Time Series. IEEE International Conference on Data Mining.

Keogh, E. and P. Smyth, (1997). A probabilistic approach to fast pattern matching in time series databases. In Proceedings of the 3rd Conference on Knowledge Discovery in Databases and Data Mining, Newport Beach, VA, USA.

Pavlidis, T. and S. Horowitz, Segmentation of Plane Curves, IEEE Transactions on Computers, vol. c-23, no. 8, August 1974.

Press, W. H., S. A. Teukolsky, W. T. Vetterling and B. P. Flannery, (1992). Numerical Recipes in C: The Art of Scientific Computing, Chap. 13, 2nd Edition, Cambridge University Press.