A ratio test for the accurate automatic P - wave onset detection

ERJON - VASILIS M. PIKOULIS *and* EMMANOUIL Z. PSARAKIS

Computer Engineering and Informatics Department, University of Patras, Rio, 26504, Greece
email: {pikoulis, psarakis}@ceid.upatras.gr

***Abstract*—In this paper a new nonlinear transformation of the raw data based on the notion of the length of seismogram and a ratio test statistic for the accurate automatic P - wave onset detection problem are proposed. The basic characteristic of the proposed statistic is that it makes no assumptions for the properties of the noise, apart from very loose stationarity requirements. Based on this statistic, we propose an algorithm for the accurate estimation of the arrival times. From a series of experiments on real signals we have conducted, the proposed picking method seems to outperform well known in the literature methods.**

I. Introduction

The determination of the arrival time of a seismic wave to a particular recording station is referred to as wave picking and automated procedures that address this problem as automatic pickers. As some of the most fundamental problems in Seismology, including event location, event identification, source mechanism analysis, relocation procedures and tomography, rely on travel - time inversion techniques, the reliability of the solutions depends heavily on the accuracy of the estimated arrival times of the waves to a network of seismic stations. Moreover, the last two of the aforementioned problems require the detection and analysis of a very large number of small magnitude events implying that the automatic picking technique needs to be both robust, as microseisms produce signals with low SNR and computationally efficient.

The first attempts to the solution of the automatic picking problem, were based on the ratio of a Short Term Average (STA) and a Long Term Average (LTA) of some Characteristic Function (CF) of the data. The general idea is that in areas of noise the value of the ratio should remain substantially constant, while when a signal emerges, the STA should be able to capture the change much more quickly than the LTA, resulting in a sudden rise of the ratio values. The decision for an arrival is then based on the comparison of the STA/LTA ratio to a - mostly empirically - predetermined threshold. The CF proposed by Allen [1], [2], is given as a weighted sum of the squared amplitude and the squared derivative of the signal, while Baer and Kradolfer [3] presented a new CF by taking the fourth power of Allen's CF and continuously normalizing its values with the running estimation of its mean and its variance. Earle and Shearer [4] used an envelope function of the seismogram, given as the square root of the sum of the squared values and the squared Hilbert transform of the signal. Aldersons [5] presented the MannekenPix, a picking procedure that is based on the Baer - Kradolfer algorithm, but adds pre- and post-processing steps in order to improve its results. Despite their age, the STA/LTA based pickers remain among the most widely used and are included in many popular seismological software packages (e.g. Earthworm, Sac2000).

A different approach to the solution of automatic picking problem is based on the AR modeling of the seismic data [6], [7], [8], [9]. Under this framework, the seismogram is considered to be composed of two different stationary processes divided at the onset point. By considering each point of the seismogram as a candidate dividing point, an AR model is fitted on each part. This leads to a sequence of model pairs and a series of modeling error values, measured at each point by the Akaike Information Criterion (AIC). The onset time is then selected as the time point that led to the best modeling results, denoted by the minimum value of the AIC sequence. Leonard and Kenett [10] propose the use of a single AR model, which is calculated only once from the initial part of the seismogram. This model is then used for the calculation of a sequence of AIC values for the whole interval, expecting that at the onset point the statistical properties of the data will change and this will lead to greater modeling errors from that point on. Again the onset time is assumed to be pointed out by the minimum of the AIC sequence. It has been reported that the AR model - based pickers require a relatively large SNR and a sudden arrival of the wave (as opposed to an emerging arrival) in order to perform well [10], [11].

The Discrete Wavelet Transform (DWT) has also been used to detect and pick the arrival time of seismic phases. Anant and Dowla [12] applied the DWT and used polarization and amplitude information contained in the wavelet. Gendron et al [13] jointly detected and classified seismic events via Bayes theorem by using features extracted from wavelet coefficients of the records. Zhang et.al. [11] obtain a denoised form of the signal by applying soft thresholding to the DWT coefficients and then calculate an AIC - like sequence without fitting AR models, based on the variances of the signal parts before and after each candidate onset point. The minimum of this sequence gives again the selected arrival time.

Der and Shumway [14], used a modified version of the CUSUM algorithm, proposed by Inclan and Tiao [15] for the detection of multiple variance changes in time series. The authors indicate the need for pre-filtering of the seismograms in order to improve the amplitude contrast between the noise and the arrival. Nakamula et. al. [16] divide a record into equal length frames and check the local and weak stationarity of each interval using the theory of the KM_2O-Langevin equations. Their method is based on the assumption that the frames are stationary as long as they include only background noise, but the stationarity will break abruptly when a seismic signal arrives and the frames include both background noise and samples of the P-wave.

Methods based on Higher Order Statistics have also been proposed. Saragiotis et.al. [17] use a sliding window over the waveform and calculate skewness and kurtosis at ever position. They estimate the arrival time by the maximum slope of the calculated sequences, anticipating that in the neighborhoods of P - wave onset the sequences' amplitude will present local maxima, due to the radically changing statistical properties of the sample. Galiana-Merino et. al. [18] base their method on the same general idea, but perform the statistical analysis (calculation of kurtosis) on the Stationary Wavelet Domain of the signal. According to the authors this leads to a more robust estimator of the arrival time.

Finally, combination of the above described methods have also been proposed. Bai and Kenett [19] use a sliding window over the seismogram and extract a set of features based on the amplitude, the instantaneous phase and the autoregressive coefficients of each frame. The decision is then based on the STA/LTA ratio for each sequence of

feature values. AR - modeling, STA/LTA and polarization information is also combined by Diehl et.al. [20].

The remaining of this paper is organized as follows. In Section II the problem formulation is presented. In Section III a nonlinear transformation of the raw data based on the notion of the length of seismogram is defined and a new ratio test statistic for the accurate automatic P - wave onset detection problem is proposed. In addition, using this statistic, an algorithm for the accurate estimation of the arrival times of the seismic events is also proposed. In Section IV where our experimental results are presented, we compare the performance of the proposed method against two well known picking methods. Finally, Section V contains our conclusions.

II. Problem Formulation

Let us denote with x_n, $n = 0, 1, \cdots$, the record from a given station and let us also assume that during the recording interval occurred K seismic events. If we denote with s_n^k, $n = 0, 1, \ldots, N_k$, the signal produced by the k-th event and with n_k the corresponding wave arrival time, then x_n can be expressed as:

$$x_n = w_n + \sum_{k=1}^{K} s_{n-n_k}^k, \tag{1}$$

where w_n is a noise process. The problem at hand is then that of the joint estimation of the number of events, K and of the arrival times n_k. We mention here that such problems are often ill-posed and it is usually necessary to impose extra conditions in order to obtain an acceptable solution. Apart from the additive white noise, seismic signals are also contaminated with seismic noise, usually a low frequency signal which is the combined result of ground motion, ocean currents, changes in temperature and atmospheric pressure during the recording interval and other location specific factors. A typical recording of a microseismic event, exhibiting the aforementioned degradations is shown in Fig. 1. As a very common

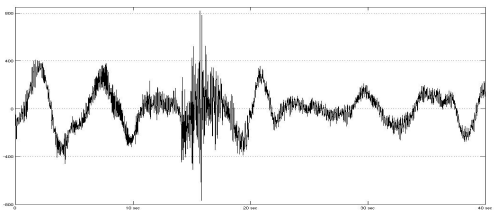


Fig. 1. Example of a seismogram.

approach to address these problems, Seismologists apply bandpass filtering to the data in a pre-processing step [14], considering that any contribution to very low and high frequencies is mainly due to noise. However, by taking into account that the most useful information for solving the problem at hand is contained in the high frequency band of the signal, the validity of the above mentioned assumptions is questionable. Indeed, we expect that the most radical change of the amplitude of the recorded signal, occurring at the arrival time of the event, will manifest itself in the high frequency content of the signal. Thus, smoothing out - basically in an uncontrollable way - the information concentrated around the time instances of the highest interest, seems to be a non convincing action. In the next section we propose a more natural quantity that emphasizes on the above mentioned point, and use it for the definition of a test statistic, in order to attack the estimation problem at hand.

III. The proposed length based test statistic

Let us consider a discrete time signal x_n obtained from the sampling of its continuous time counterpart $x(t)$ with a sampling period of T_s, so that $x_n = x(nT_s)$. Let us also define $\Delta\mathcal{L}_n,\ n = 1, 2, \cdots$ as the Euclidean length of the line segment connecting consecutive pairs of the points $((n-1)T_s, x_{n-1})$ and (nT_s, x_n), i.e.:

$$\Delta\mathcal{L}_n = \left((x_n - x_{n-1})^2 + T_s^2\right)^{\frac{1}{2}} = T_s \left((\frac{x_n - x_{n-1}}{T_s})^2 + 1 \right)^{\frac{1}{2}}. \tag{2}$$

In order to give a more physical meaning in the above defined quantity, let us concentrate ourselves in the noiseless case, i.e. $w_n = 0$. If we consider that the first order backward differences of signal x_n appeared in Equ. (2) constitute an approximation[1] of the derivative $\dot{x}(t)$ of function $x(t)$ at the sampling point nT_s, i.e.:

$$\frac{x_n - x_{n-1}}{T_s} \approx \dot{x}(t)|_{t=nT_s}, \tag{3}$$

then $\Delta\mathcal{L}_n/T_s$ can be considered as the approximation of the instantaneous change of the length of curve (let us denote it by $\mathcal{C}(x)$), defined by the following relation:

$$\dot{\mathcal{L}}(t) = (\dot{x}^2(t) + 1)^{\frac{1}{2}}, \tag{4}$$

at the same points, i.e.:

$$\frac{\Delta\mathcal{L}_n}{T_s} \approx \dot{\mathcal{L}}(t)|_{t=nT_s}. \tag{5}$$

Note that Equ. (4) expresses the instantaneous change in the length of $\mathcal{C}(x)$ as a function of the first derivative of $x(t)$. Note also that the quantity defined in Equ. (2) can also be considered as a highly nonlinearly filtered version of the original signal, with its high frequency content enhanced and at the same time, its low frequencies suppressed, thus ensuring, in some sense, the requirements mentioned in the last paragraph of the Section II.

However, in the presence of noise $x(t)$ and consequently its sampled counterpart x_n are stochastic processes, meaning that the values of $\Delta\mathcal{L}_n$ are in fact random variables (RVs), the statistical properties of which can only be derived under particular assumptions for the properties of the noise. Note however that in the case of the seismic noise these properties are not only unknown, but it also not safe to infer them as the factors governing the behavior of the noise are not so clear. For this reason we would like the proposed transformation of the raw data to ensure that our judge will be based only on very mild stationarity requirements of the noise process. As we can see from Fig. 2 where the evolution of the values of $\Delta\mathcal{L}_n$ with time for the record of Fig.1 is shown, the proposed length based quantity seems to ensure this requirement.

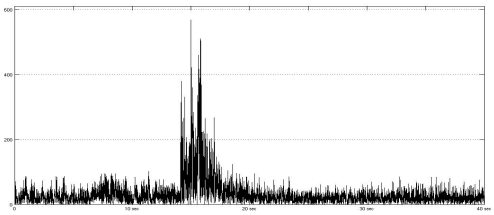


Fig. 2. Values of $\Delta\mathcal{L}_n$ for the seismogram of Fig.1.

In order to be able to derive a proper statistic and use it for solving the problem at hand, let us define the sequences $L_n^{N^+}$ and $L_n^{M^-}$, as the mean value of $\Delta\mathcal{L}_n$ over intervals of length NT_s and MT_s, starting and ending at the n-th sampling point respectively, i.e.:

$$L_n^{N^+} = \frac{\sum_{k=n}^{n+N-1} \Delta\mathcal{L}_k}{N}, \quad L_n^{M^-} = \frac{\sum_{k=n-M+1}^{n} \Delta\mathcal{L}_k}{M}, \tag{6}$$

[1]We assume that the real function $x(t)$ and its first derivative $\dot{x}(t) = dx(t)/dt$ are both continuous in $\mathbb{R}$.

and the following sequence of ratios:

$$\lambda_n = \frac{L_n^{N^+}}{L_{n-1}^{M^-}},\ n = 0, 1, \cdots . \tag{7}$$

This ratio constitutes the proposed test statistic for the problem at hand. Although the above defined test statistic can be used in a sequential manner for the detection of multiple events, in the following we concentrate our attention on the case of $K = 1$, where as we can see from Equ. (1), the problem is limited to the estimation of n_0. Intuitively, since $\Delta\mathcal{L}_n$ is a nonlinear function of the first order differences of x_n, and due to the particular nature of seismic noise described above, we expect the values of $L_{n-1}^{M^-}$ and $L_n^{N^+}$, to be in close vicinity of one another, as long as both time windows cover noise parts of the record. Because of this, the values of λ_n for $n \leq n_0 - N$ are expected to vary mildly around a constant level of 1. For $n_0 - N + 1 \leq n \leq n_0$, the window corresponding to $L_n^{N^+}$ will gradually cover the beginning of the seismic signal s_n, thus causing the values of $L_n^{N^+}$ to grow, while $L_{n-1}^{M^-}$ will still account only for noise. This results in a gradual rise in the values of λ_n, attaining ideally its maximum at $n = n_0$, where the whole right-hand window is placed over signal and the whole left-hand one is placed over noise. For $n > n_0$, as values of s_n start entering $L_{n-1}^{M^-}$ and as - due to the inherent fading nature of s_n - $L_n^{N^+}$ will remain almost constant, the values of λ_n will exhibit a steep drop, returning to its previous level. This behavior of λ_n, for $N = M = 50$ samples, is displayed in Fig. 3. A zoomed portion of the plot focusing on the main lobe, is displayed in the upper right corner of the same figure.

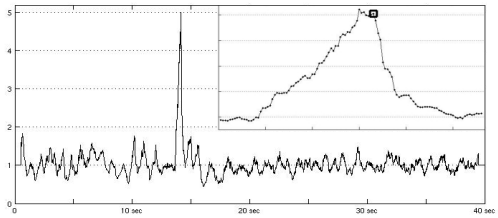


Fig. 3. Values of λ_n obtained from the $\Delta\mathcal{L}_n$ sequence shown in Fig.2.

A. Solving the estimation problem

Let us now concentrate ourselves on the solution of the desired estimation problem, namely the estimation of n_0. Following the analysis presented above, a natural selection to achieve our goal would be the solution of the following maximization problem:

$$\widehat{n}_0 = \arg\max_n \lambda_n, \tag{8}$$

i.e. the location where the sequence λ_n attains its maximum value. Clearly, the estimation of n_0 by means of $\widehat{n}_0$ is based on the assumption of a constant increase of λ_n for $n_0 - N + 1 \leq n \leq n_0$ and a constant decrease for $n_0 + 1 \leq n \leq n_0 + M - 1$, for the reasons presented in the preceding section. While this assumption is true in a noiseless case, in the presence of noise, λ_n and consequently $\widehat{n}_0$ are random variables. As a result, the general validity of the above mentioned assumption is influenced by factors such as the strength of noise near the arrival time, the amplitude and the shape of the signal as well as the window sizes N and M used in the calculation of λ_n. A particularly non favorable situation occurs in the case of an impulsive arrival followed by a quick decay of the signal. In such a case, the change of the signal curve length $\Delta\mathcal{L}_n$ will be considerably higher than that corresponding to noise samples, only in a limited interval following n_0. Denoting the length of this interval by P, and assuming that $N > P$, the values of $L_n^{N^+}$ will increase only in the interval $n_0 - N + 1 \leq n \leq n_0 - N + P$, remaining virtually constant in the interval $n_0 - N + P + 1 \leq n \leq n_0$. Thus, in the latter interval the value of λ_n will be mainly affected by the values of the denominator of the ratio defined in Equ. (7), which in turn depend on the characteristics of the noise (e.g. stationarity in mean and variance) and the size M of the window. This results in an uncertainty interval of $N - P$ samples where the location of the true maximum value of the sequence is basically unpredictable. As a consequence, the optimum solution of (8) is sensitive to all the factors mentioned above, thus degrading the performance of $\widehat{n}_0$ as an estimator of n_0, leading to estimates that point out to a time, a few samples prior to the true arrival.

Instead of relying on the maximum value of λ_n, a different approach to the solution of the problem at hand, exploits the fact that the values of λ_n start to drop rapidly, as soon as the signal samples start entering its denominator. This is an inherent feature of λ_n which is more insensitive to the parameters affecting the location of the maximum. Thus, determining the point of the beginning of the steep drop of λ_n, or schematically the rightmost corner of the maximal peak of λ_n, results in a more consistent estimator of n_0. For the determination of this point we propose the following two step procedure:

S_1: Find a point located on the descending slope of the maximal peak by solving the following maximization problem:

$$\widetilde{n}_0 = \arg\max_n \lambda_n(\lambda_{n-1} - \lambda_n). \tag{9}$$

We anticipate the solution of (9) to be a point that combines a large λ_n with a significant decrease in value, exhibiting therefore the desired characteristics.

S_2: "Climb" the slope until the corner is reached (**while** $\lambda_{\widetilde{n}_0} - \lambda_{\widetilde{n}_0 - 1} < 0$ **do** $\widetilde{n}_0 = \widetilde{n}_0 - 1$).

Note that in cases that are favorable for $\widehat{n}_0$, the two estimates of n_0 will coincide. The event of Fig. 1 does not represent such a case, as can be seen in the upper right corner of Fig. 3, where the estimation returned by $\widetilde{n}_0$ is marked by the small square, whereas the one returned by $\widehat{n}_0$ can be discerned a few samples earlier. This is not by accident, as we are going to see in the experimental results, presented in the next section. In Fig.4, the two estimates are displayed against the recorded signal for a more comprehensive view (the value of $\widehat{n}_0$ is shown by the solid line, and the value of $\widetilde{n}_0$, which corresponds to the true value of n_0 for this case, by the dashed line).

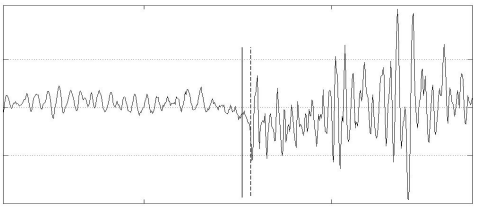
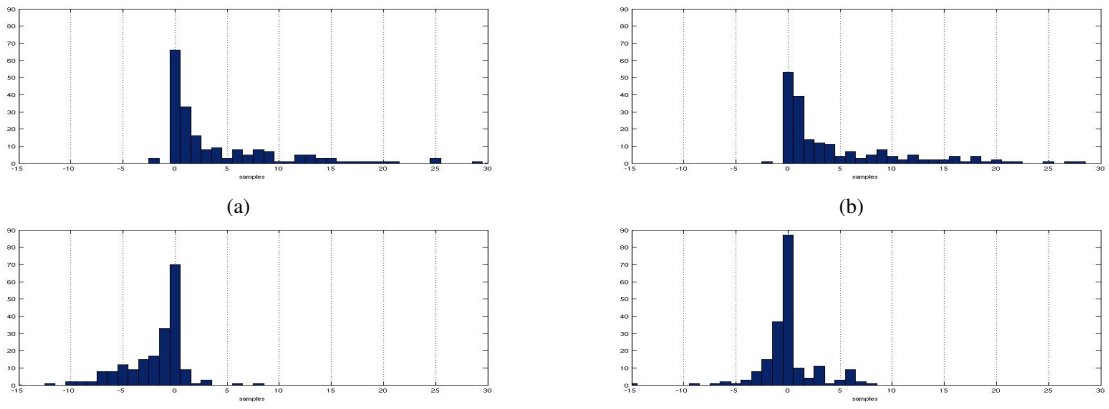


Fig. 4. Estimation of n_0 from the values of λ_n shown in Fig.3.

IV. Experimental Results

In this section we test the accuracy of the proposed picking method (both estimators) against the methods proposed in [14] (CUSUM) and [17] (HOS). In order to achieve our goal, we used a data set of 200 pre-cut recordings, each containing one seismic event. The results obtained by each method were compared to the "true" arrival times, manually picked by a human analyst and the histograms of the picking errors are displayed in a unified scale in Fig. 5. For the sake of fairness, all the methods were applied to the length sequences ($\Delta\mathcal{L}_n$) obtained from the raw data. Experiments on bandpass filtered versions



(a) (b) (c) (d)

Fig. 5. Histograms of picking errors obtained by the application of the methods under comparison. (a): CUSUM based method. (b): HOS based method. (c): Proposed method based on the solution of the optimization problem (8). (d) Proposed method based on the two step procedure.

of data were also conducted, but due to lack of space, they are not presented. However, we must stress at this point that the results we obtained from the application of all methods, were significantly worse than the length-based ones thus revealing, in some sense, the appropriateness of the proposed transformation of the raw data. As we can clearly see from Fig. 5 (a-b), both the CUSUM and the HOS based methods, exhibit quite similar performance, with their picking errors asymmetrically distributed towards positive values, indicating a constant delay in the detection of the arrived signal. For the CUSUM method the mean error was 4.5 samples with a standard deviation (std) of 7 samples, while for the HOS based method the mean error was 4.6 samples with a std of 6.2 samples. On the other hand, as it is clearly depicted in Fig. 5 (c-d) the proposed estimators have a larger percentage of error-free pickings and their distributions are more symmetrical than the ones achieved by their rivals. Especially, the histogram shown in Fig. 5 (c) is almost symmetrical with respect to 0, thus revealing its superiority among all the other ones. Regarding the performance of the proposed estimators, $\widehat{n}_0$ resulted in a mean error of -2.2 samples, with a std of 3.9 samples, and $\widetilde{n}_0$ returned a mean error of -0.02 samples, with a std of 2.65 samples.

V. Conclusions

In this paper a ratio test statistic based on the use of a length based quantity tailored to the estimation problem of the seismic event arrival time, and two estimators for its solution were proposed. The performance of the proposed estimators were compared against two well known picking ones. The experimental results confirm that the proposed methods outperform their rivals. Issues concerning the generalization of the proposed methods for their applicability to the case of multiple seismic events, are currently under investigation.

Acknowledgments

The authors would like to thank the Seismological Laboratory of University of Patras, for their support in providing the experimental data set and for offering their expertise on several seismological issues.

This work was financed by the University of Patras, Karatheodori research program, entitled “The relocation problem of seismic event hypocenter parameters”.

References

[1] R.V. Allen, *Automatic earthquake recognition and timing from single traces*, Bull. Seism. Soc. Am. 68, 1521-1532, 1978.

[2] R.V. Allen, *Automatic phase pickers: their present and future prospects*, Bull. Seism. Soc. Am. 68, S225-S242, 1982.

[3] M. Baer, and U. Kradolfer, *An Automatic phase picker for local and teleseismic events*, Bull. Seism. Soc. Am. 77, 1437-1445, 1987.

[4] P. Earle, and P. Shearer, *Characterization of global seismograms using an automatic picking algorithm*, Bull. Seism. Soc. Am. 84, no. 2, 366-376, 1994.

[5] F. Aldersons, *Toward a three-dimensional crustal structure of the Dead Sea region from local earthquake tomography*, PhD thesis, Tel Aviv University, Israel, 2004.

[6] A. Kushnir, V. Lapshin, V. Pinsky, and J. Fyen, *Statistically optimal event detection using small array data*, Bull. Seism. Soc. Am. 80, no. 6b, 1934-1950, 1990.

[7] T. Takanami, and G. Kitagawa, *A new efficient procedure for the estimation of onset times of seismic waves* J. Phys. Earth 36, 267-290, 1988.

[8] T. Takanami, and G. Kitagawa, *Estimation of the arrival times of seismic waves by multivariate time series models* Ann. Inst. Stat. Math. 43 (3), 407-433, 1991.

[9] T. Kværna, *Automatic onset time estimation based on autoregressive processing*, Norsar scientific report, NORSAR, 1995.

[10] M. Leonard, and B.L.N. Kennett, *Multi-component autoregressive techniques for the analysis of seismograms*, Phys. Earth Planet. Int. 113, no. 2, 247-264, 1999.

[11] H. Zhang, C. Thurber, and C. Rowe, *Automatic P-Wave Arrival Detection and Picking with Multiscale Wavelet Analysis for Single-Component Recordings*, Bull. Seism. Soc. Am. 93, no. 5, 1904-1912, 2003.

[12] S.K Anant, and F.U. Dowla, *Wavelet transform methods for phase identification in three-component seismogram*, Bull. Seism. Soc. Am. 87, 1598-1612, 1997.

[13] P. Gendron, J. Ebel, and D. Manolakis, *Rapid joint detection and classification with wavelet bases via Bayes theorem*, Bull. Seism. Soc. Am. 90, 764-774, 2000.

[14] Z.A. Der, and R.H. Shumway, *Phase onset time estimation at regional distances using the CUSUM algorithm*, Phys. Earth Planet. Int. 113, no. 2, 227-246, 1999.

[15] C. Inclan, and G.C. Tiao, *Use of Cumulative Sums of Squares for Retrospective Detection of Changes of Variance*, J. Amer. Statist. Assoc. 89, 913-923, 1994.

[16] S. Nakamula, M. Takeo, Y. Okabe, and M. Matsuura, *Automatic seismic wave arrival detection and picking with stationary analysis: Application of the KM_2O - Langevin equations*, Earth Planets Space 59, 567-577, 2007.

[17] C.D. Saragiotis, L.J. Hadjileontiadis, and S.M. Panas, *PAI-S/K: A robust automatic seismic P phase arrival identification scheme*, IEEE Trans. Geosci. Remote Sens. 40, 1395-1404, 2002.

[18] J.J. Galiana-Merino, J.L. Rosa-Herranz, and S. Parolai *Seismic P Phase Picking Using a Kurtosis-Based Criterion in the Stationary Wavelet Domain*, IEEE Trans. Geosci. Remote Sens. 46, 3815-3825, 2008.

[19] C. Bai, and B.L.N. Kennett, *Automatic Phase-Detection and Identification by Full Use of a Single Three-Component Broadband Seismogram*, Bull. Seism. Soc. Am. 90, 187-198, 2000.

[20] T. Diehl, N. Deichmann, E. Kissling, and S. Husen, *Automatic S-Wave Picker for Local Earthquake Tomography*, Bull. Seism. Soc. Am. 99, no. 3, 1906-1920, 2009.