AN APPROACH FOR DENOISING WAVEFORM DATA BY AUTO-REGRESSIVE ALGORITHM

Nobuo Arai, Takahiko Murayama, Makiko Iwakuni, and Mami Nogami

National Data Center-1 of Japan (Japan Weather Association)

Sponsored by the CTBT National Operations, Ministry of Foreign Affairs, Japan

ABSTRACT

The waveform data from the Comprehensive Nuclear-Test-Ban Treaty (CTBT) International Monitoring System (IMS) are mainly used for making the epicenter and characteristics of an observed event clear. Waveform data usually consist of signals from natural or man-made phenomena and site-specific noise from the ambient surroundings.

Unfortunately, if the dominant frequency of a signal is within the frequency band of the site-specific noise and the amplitude of the signal is close to the detection limit, the signal may be buried within the noise.

Bandpass filtering of data is often used to reduce the influence of the background noise (such as the site-specific noise) and enhance signal-to-noise ratios (SNRs). However, this approach is not adequate for extracting signals from raw data because such an approach does not distinguish the nature of different waveforms.

In order to remove site-specific noise components from observed raw data, an auto-regressive (AR) algorithm can be applied to extract a pure signal. The time series of the background noise is simulated by an AR model, and then a pure signal would be extracted by removing the simulated noise component from the raw data.

Effectiveness in utilizing such a concept has been tested on infrasound data observed at the IMS with affirmative results.

In this research, we demonstrate both the possibility and the capability of denoising waveform data utilizing the AR algorithm technique.

OBJECTIVES

Sources of infrasonic noise predominantly consist of eddies generated by local atmospheric disturbances (e.g., "local winds"). Reducing the influence of local winds or extracting a signal from noisy data is a significant issue at many CTBT IMS infrasound monitoring stations.

We present data from the CTBT infrasound monitoring station, IS30. Although each array element is designed to minimize the effects of wind-generated noise, it is difficult to extract signals from observed data when strong winds occur near the station.

Bandpass filtering of data is often used to remove this wind-generated noise. However, this approach is not adequate for extracting signals from raw data because such an approach does not distinguish the nature of different waveforms. Moreover, if the dominant frequency of the signal is within the frequency band of the site-specific noise and the amplitude of the signal is close to the detection limit, the signal may be buried within the noise.

Therefore, the objective of this research is to reduce the influence of background noise (such as wind-generated noise) and enhance the SNR by using a new technique that is frequency-band independent.

RESEARCH CONDUCTED

Infrasonic noise is predominantly wind-generated. Generation of wind noise is described as a stochastic process, and the state of wind can be stationary within a short period.

Thus, to remove site-specific noise components from observed raw data, an AR algorithm is applied for the extraction of a pure signal. The time series of the background noise is simulated by an AR model, and the pure signal is extracted by removing the simulated noise from the observed raw data.

The AR algorithm is a new technique for denoising waveform data and consists of three components: the simulation method of the background noise, the reduction procedure of the background noise, and the extraction of pure signal, and is described as follows.

Fundamental conceptual model of the waveform data

The concepts of denoising and the subsequent extraction of pure signal are shown in Figure 1.

As shown schematically in the outline below, the approach we used in this research consisted of eliminating background noise from the observed data so as to obtain pure signal.

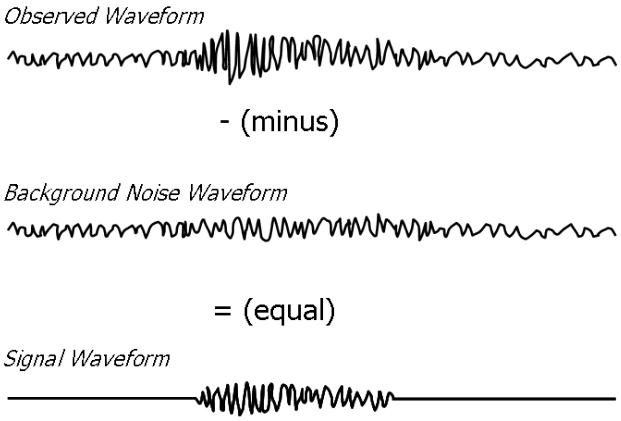

Figure 1. Conceptual image of denoising procedure on waveform data.

The actual observed infrasound data (y_n) also contains a trend component in addition to the background noise illustrated in Figure 1.

Hence, we assumed the following equation based on the time series analysis method (e.g., Vaseghi, 1996; Kitagawa, 1993):

$$y_n = t_n + r_n + s_n + w_n \quad (1)$$

where t_n is the trend component, r_n is background noise, s_n is signal, and w_n is observation noise, at time n.

State space model

In order to estimate the time series of each component, we utilize the state space model. With the assumption that y_n is the time series of l variate, the state space model is represented by the following equation:

$$x_n = Fx_{n-1} + Gw_n \quad \text{(System Model)} \quad (2)$$

$$y_n = Hx_n + \varepsilon_n \quad \text{(Observation Model)} \quad (3)$$

where x_n is the state vector, which is directly unobservable; k is the dimensional vector; w_n is the system noise (or the state noise); and m is the order dimensional white noise in accordance with zero mean and the variance-covariance matrix (Q_n).

On the other hand, ε_n is the observation noise, and l is the order dimensional white noise in accordance with zero mean and the variance-covariance matrix (R_n).

F, G, and H are matrixes of $k \times k$, $k \times m$, and $l \times k$, respectively.

The state space model is interpreted in two ways, as follows:

1) The observation model (Equation [3]) is assumed to be the regressive model.

   In this case, the model shows the mechanism in which y_n is observed. The state x_n shows the regression coefficient, and the system model shows the change of the regression coefficient with respect to time.

2) The state vector x_n is the signal to be estimated.

   In this case, the system model shows the mechanism of the signal generation, and the observation model shows the shape when a signal is actually observed, with the noise component added to that signal.

Applied concrete cases are explained in terms of each component model.

The following section describes each component model (r_n, background noise; s_n, signal; and t_n, trend component;).

Background noise component model and signal component model

We have modeled the background noise component r_n and the signal component s_n using an AR model because their time series' are short period in nature.

An AR model is a type of random process that is often used to model and predict various types of natural phenomena. AR models of background noise r_n and signal s_n are given by the following equations:

$$r_n = \sum_{j=1}^{m} a_j r_{n-j} + u_n \quad (4)$$

$$s_n = \sum_{j=1}^{l} b_j s_{n-j} + v_{n1} \quad (5)$$

These are based on parameters a_j and b_j, where $j = 1,...,m$ and $j = 1,...,l$. There is a direct correspondence between these parameters and the covariance function of the process. This correspondence can be inverted to determine the parameters from the autocorrelation function (which is itself obtained from the covariances), where u_n and v_{n1} are white noise processes each with a zero mean and respective variances τ_1^2, τ_2^2.

Using Equation (4) as an example and defining the state vector as $x_n = (r_n, r_{n-1}, \cdots, r_{n-m+1})^t$, the following equations hold between x_n and x_{n-1}:

$$x_n = F_1 x_{n-1} + G_1 u_n \tag{6}$$

$$x_n = \begin{bmatrix} r_n \\ r_{n-1} \\ \vdots \\ r_{n-m+1} \end{bmatrix}, \quad F_1 = \begin{bmatrix} a_1 & a_2 & \cdots & a_m \\ 1 & & & \\ & \ddots & & \\ & & 1 & 0 \end{bmatrix}, \quad G_1 = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \tag{7}$$

where F_1 and G_1 is a matrix of $m \times m$ and an m-dimensional vector, respectively. And, as the first component of the state (x_n) is r_n, the observation model (r_n) (Equation [8]) below, given $H_1 = \begin{bmatrix} 1 & 0 & \cdots & 0 \end{bmatrix}$ is:

$$r_n = H_1 x_n \tag{8}$$

The variances of the system noise and observation noise are $Q = \tau_1^2$ and $R = 0$; thus, the state space model of the AR model is given by Equation (8).

As in the case of the background noise model, when the matrices of the state space model are assumed, see Equations (9) and (10), the state space model of the signal component is represented by Equation (11).

$$x_n = \begin{bmatrix} s_n \\ s_{n-1} \\ \vdots \\ s_{n-m+1} \end{bmatrix}, \quad F_2 = \begin{bmatrix} b_1 & b_2 & \cdots & b_m \\ 1 & & & \\ & \ddots & & \\ & & 1 & 0 \end{bmatrix}, \quad G_2 = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \quad H_2 = \begin{bmatrix} 1 & 0 & \cdots & 0 \end{bmatrix} \tag{9}$$

$$x_n = F_2 x_{n-1} + G_2 v_{n1} \tag{10}$$

$$s_n = H_2 x_n \tag{11}$$

Trend component model

Infrasound observations contain long-period pressure changes (from a few minutes to a dozen or so minutes), such as a gravity wave of the Earth. We estimated the time series of a long-period wave by using the trend component model t_n.

The simplest way to estimate the trend of a time series is to use the polynomial trend model. This model is expressed by the following equation (the time series [y_n] is composed of the polynomial equation [t_n] and residuals [w_n]):

$$y_n = t_n + w_n \tag{12}$$

where w_n is a white noise process with zero mean and variance σ^2. The trend component model t_n is the polynomial equation represented by the following equation:

$$t_n = a_0 + a_1 x_n + \cdots + a_m x_n^m \tag{13}$$

This equation is the specific model. In order to extend this model to a more flexible function, the k^{th}-stochastic difference equation (14) is provided as a useful alternative. In this equation, it is assumed that the k^{th} difference of t_n is close to zero.

$$\Delta^k t_n = v_{n2} \tag{14}$$

where v_{n2} is a white noise process with zero mean and variance τ_3^2.

The trend component model (14) is generally represented by the following equation, given $\Delta \equiv 1 - B$ (the time subtraction operator defined by $\Delta t_n \equiv t_n - t_{n-1}$):

$$\Delta^k = (1 - B)^k = \sum_{i=0}^{k} {}_k c_i (-B)^i \tag{15}$$

When c_k is defined by $c_i = (-1)^{i+1}{}_k c_i$, Equation (15) is represented by the following:

$$t_n = \sum_{j=1}^{k} c_k t_{n-i} + v_{n2} \tag{16}$$

This model can be considered the AR model of the k^{th}-order dimension although not the stationary state. Therefore, when the matrices of the state space model are assumed, as in Equations (17) and (18), the state space model of the trend component model is represented by Equation (19).

$$x_n = \begin{bmatrix} t_n \\ t_{n-1} \\ \vdots \\ t_{n-k+1} \end{bmatrix}, \quad F_3 = \begin{bmatrix} c_1 & c_2 & \cdots & c_k \\ 1 & & & \\ & \ddots & & \\ & & 1 & 0 \end{bmatrix}, \quad G_3 = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \quad H_3 = \begin{bmatrix} 1 & 0 & \cdots & 0 \end{bmatrix} \tag{17}$$

$$x_n = F_3 x_{n-1} + G_3 v_n \tag{18}$$

$$t_n = H_3 x_n \tag{19}$$

The parameter setting of the state space model for the resolution of the observed infrasound data

When the equations for each time series component are combined with the state space model explained above, the following parameters of the state space model are provided.

$$F = \begin{bmatrix} F_1 & & \\ & F_2 & \\ & & F_3 \end{bmatrix}, \quad G = \begin{bmatrix} G_1 & & \\ & G_2 & \\ & & G_3 \end{bmatrix}, \quad H = \begin{bmatrix} H_1 & H_2 & H_3 \end{bmatrix}, \quad Q_n = \begin{bmatrix} \tau_1^2 & 0 & 0 \\ 0 & \tau_2^2 & 0 \\ 0 & 0 & \tau_3^2 \end{bmatrix} \tag{20}$$

By using the state space model defined by Equation (20), the observed data are broken down into each waveform (the trend, the background noise, and the signal).

Thus, we can take away the trend and background noise from the observed data to finally get a waveform consisting of pure signal.

Kalman filter

In dealing with the state space model, it is a vital hypothesis that the state (x_n) is estimated based on the observed data of the time series (y_n). In order to estimate this state, we have used the Kalman filter algorithm in this research. The Kalman filter is a recursive estimator and based on linear dynamical systems discretized in the time domain. We could effectively calculate the conditional marginal distribution of the state (x_n). In what follows, the notation $x_{n/m}$ represents the estimate of x at time n, given observations up to and including time m.

The state of the filter is represented by two variables:

$x_{n/m}$: the estimate of the state at time n given observations up to and including time m

$V_{n/m}$: the error covariance matrix (a measure of the estimated accuracy of the state estimate)

The Kalman filter has two distinct phases: <Predict> and <Update>. The predict phase uses the state estimate from the previous time-step to produce an estimate of the state at the current time-step. In the update phase, measurement information at the current time-step is used to refine the prediction to arrive at a new, more accurate state estimate, again for the current time-step.

< Predict >

$$x_{n/m} = F_n x_{m/m} \quad \text{(Predicted state)} \quad (21)$$

$$V_{n/m} = F_n V_{m/m} F_n^{\ t} + G_n Q_n G_n^{\ t} \quad \text{(Predicted estimate covariance)} \quad (22)$$

< Update >

$$K_n = V_{n/m} H_n^{\ t} (H_n V_{n/m} H_n^{\ t} + R_n)^{-1} \quad \text{(Kalman gain)} \quad (23)$$

$$x_{n/n} = x_{n/m} + K_n (y_n - H_n x_{n/m}) \quad \text{(Updated state estimate)} \quad (24)$$

$$V_{n/n} = (I - K_n H_n) V_{n/m} \quad \text{(Updated estimate covariance)} \quad (25)$$

Order dimension of the trend model and AR model

With regards to the order dimension of the trend model and AR model, we have used the Akaike Information Criterion (AIC). AIC is a measure of how well an estimated statistical model fits (e.g., Sakamoto, etc., 1981). It is based on the concept of entropy, which in effect offers a relative measure of the information lost when a given model is used to describe reality and can be said to describe the tradeoff between bias and variance in model construction, or loosely speaking, that of precision and complexity of the model.

AIC is a tool for model selection. Given a dataset, several competing models may be ranked according to their AIC, with the one having the lowest AIC being the best. From the AIC value, one may infer that, for example, the top three models are in a tie and the rest are far worse, but one should not assign a value above which a given model is "rejected."

The AIC is derived by the following equation:

$$AIC = 2k - 2\ell(\hat{\theta}) \quad (26)$$

where k is the number of parameters in the statistical model, and $\hat{\theta}$ is the maximized value of the log-likelihood function for the estimated model.

Examples of the Application of Waveform Denoising

Examples of removing the trend and the noise, resulting in the extraction of pure signal, are shown in Figures 2 through 4.

Below is a case of an aircraft that crashed while attempting to land at the Narita Airport in Japan on March 22, 2009. The infrasound signal was generated by the explosion, and CTBT IMS infrasound monitoring station IS30 detected the signal. The figure below shows that once the trend and the noise components are eliminated, pure signal is clearly detected.

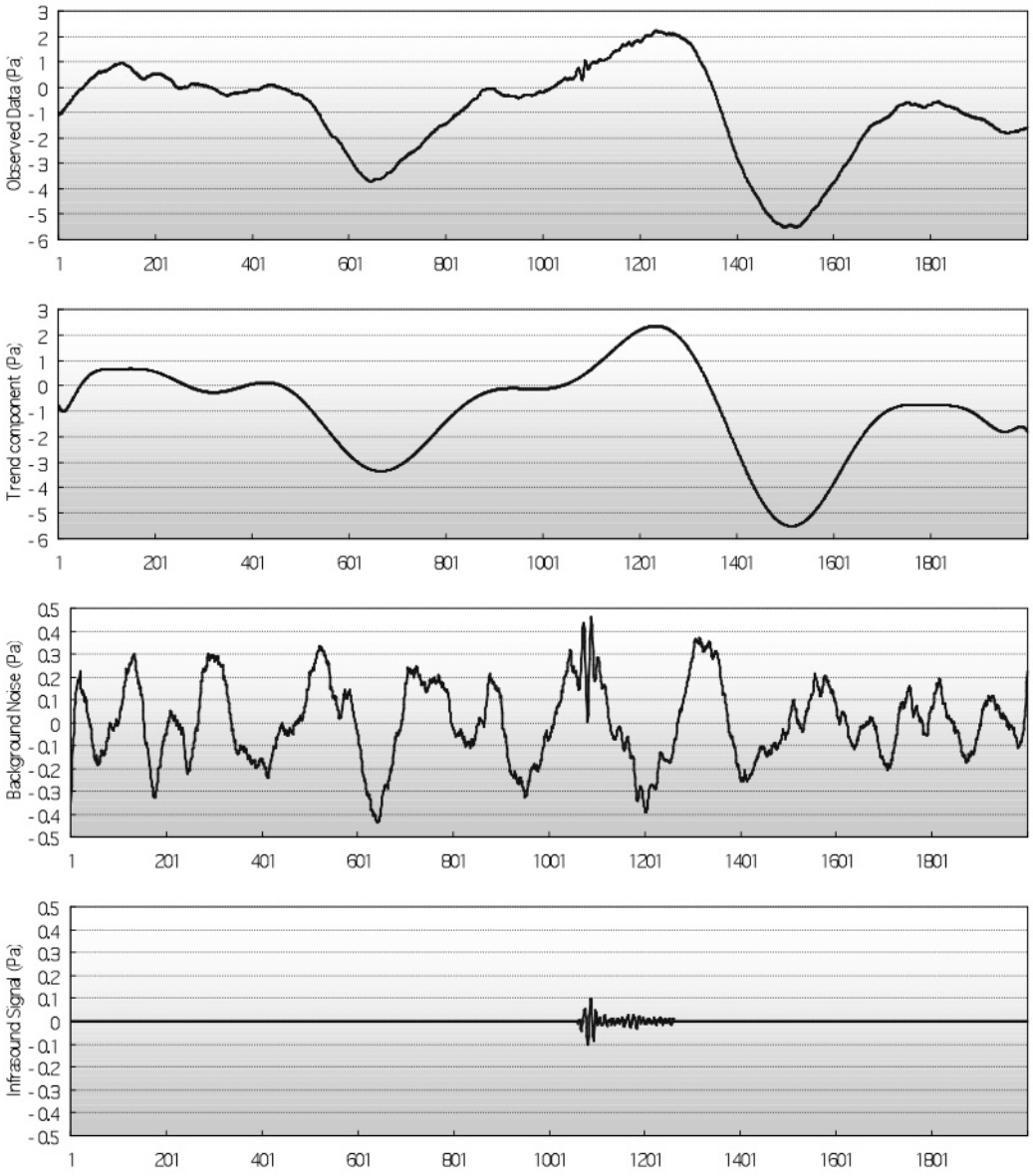

Figure 2. The infrasound waveform (documented during an aircraft accident on 22 March) recorded at the CTBT IMS infrasound station IS30 and the estimated results of each time series, namely (top) the observed raw data, (top middle) the trend component simulated by the trend component model, (bottom middle) the background noise simulated by the AR model, and (bottom) the extracted infrasound signal.

The case of a volcanic eruption is shown in Figure 3.

Mt. Asama, which is located in central Japan, experienced a minor eruption on February 2, 2009, and the CTBT IMS infrasound monitoring station IS30 recorded an infrasound signal. By means of eliminating the trend and the noise components from the observed data, the infrasound signal generated by the eruption is clearly detected.

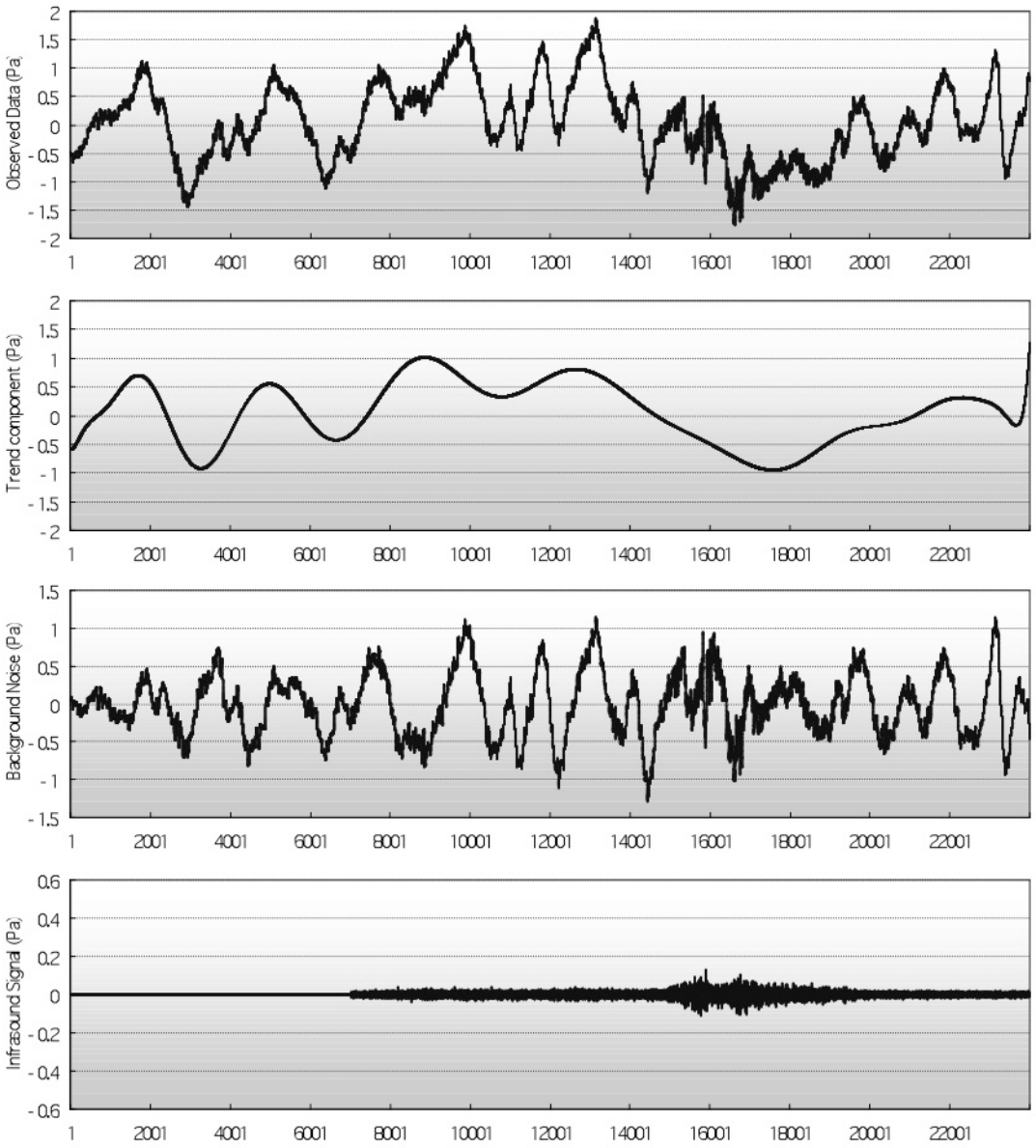

Figure 3. The infrasound waveform (produced as a result of the Mt. Asama eruption on February 2, 2009) recorded at the CTBT IMS infrasound station IS30 and the estimated results of each time series, namely (top) the observed raw data, (top middle) the trend component simulated by the trend component model, (bottom middle) the background noise simulated by the AR model, and (bottom) the extracted infrasound signal.

A lightning occurrence is shown in Figure 4.

IS30 recorded an infrasound signal when lightning struck approximately 60 km away from the array on July 27, 2008. After eliminating the trend and the noise components, the pulsed infrasound signal produced by the lightning is detected.

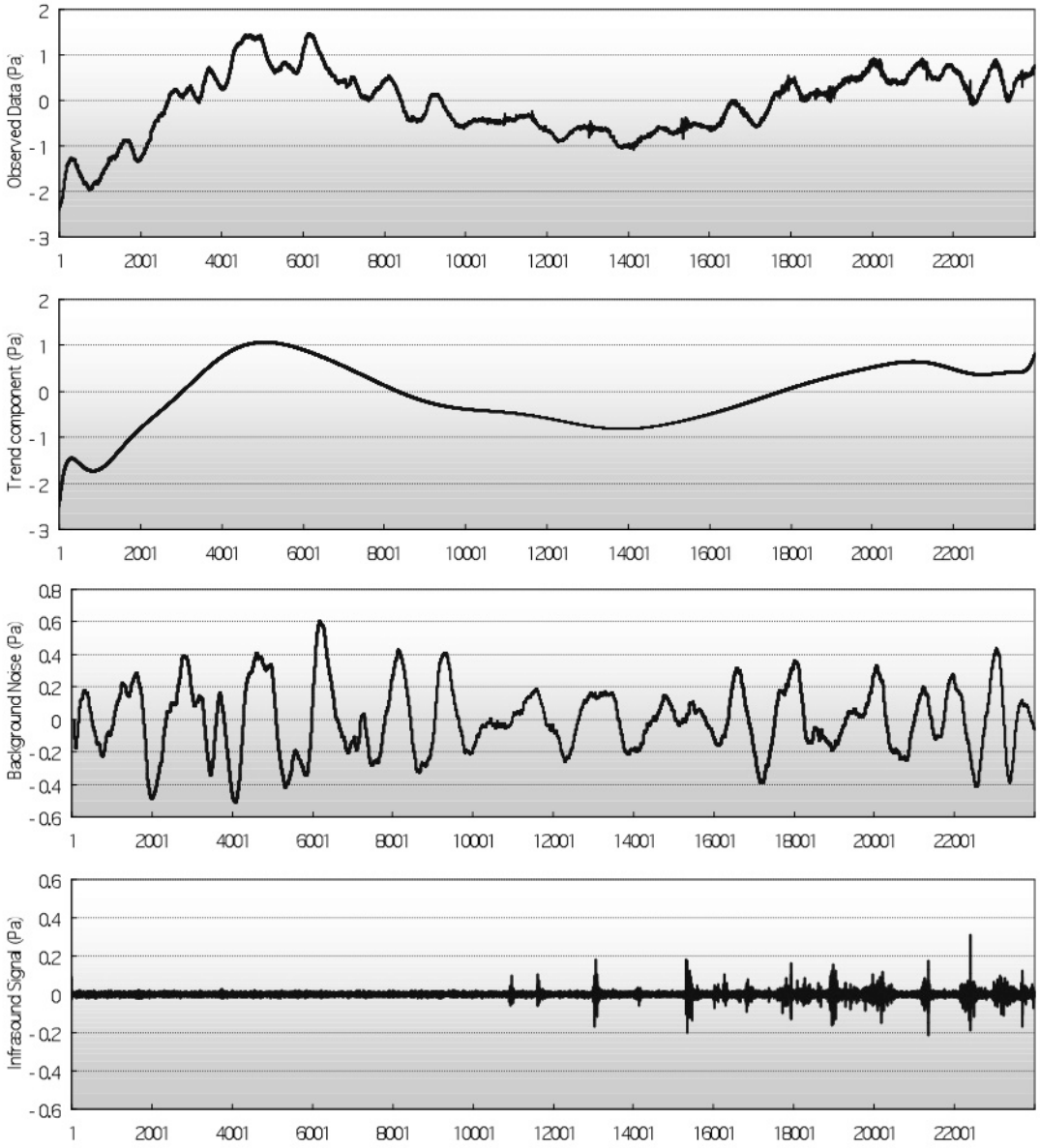

Figure 4. The infrasound waveform (produced by lightning on July 27, 2008) recorded at the CTBT IMS infrasound station IS30 and the estimated results of each time series, namely (top) the observed raw data, (top middle) the trend component simulated by the trend component model, (bottom middle) the background noise simulated by the AR model, and (bottom) the extracted infrasound signal.

CONCLUSIONS AND RECOMMENDATIONS

As has been demonstrated above, in order to remove site-specific noise (such as wind-generated noise) components from observed raw data, an AR algorithm can be applied to extract a pure signal. The time series of the trend and the noise have been simulated by the trend model and AR model respectively, and then a pure signal can be extracted by removing the simulated noise component from the observed raw data.

Effectiveness in utilizing such a concept has been tested by infrasound data observed at the CTBT IMS infrasound monitoring station IS30, with positive results in several cases. However, the effectiveness of the AR model was not substantially discussed and examined and should be continuously evaluated by using other infrasound monitoring data.

REFERENCES

Vaseghi, S. (1996). Advanced digital signal processing and noise reduction, Second edition, John Wiley & Sons Ltd.

Kitagawa, G. (1993). The time series analysis programming, Iwanami Shoten.

Sakamoto, Y., etc. (1981). Akaike Information Criterion, Statistics, D. Reidel Publishing Company.