International Journal of Chemical, Materials and Biomolecular Sciences
ISSN: 2415-6620
Vol:2, No:7, 2008

Issues in Spectral Source Separation Techniques for Plant-wide Oscillation Detection and Diagnosis

A.K. Tangirala and S. Babji

***Abstract*—** In the last few years, three multivariate spectral analysis techniques namely, Principal Component Analysis (PCA), Independent Component Analysis (ICA) and Non-negative Matrix Factorization (NMF) have emerged as effective tools for oscillation detection and isolation. While the first method is used in determining the number of oscillatory sources, the latter two methods are used to identify source signatures by formulating the detection problem as a source identification problem in the spectral domain. In this paper, we present a critical drawback of the underlying linear (mixing) model which strongly limits the ability of the associated source separation methods to determine the number of sources and/or identify the physical source signatures. It is shown that the assumed mixing model is only valid if each unit of the process gives equal weighting (all-pass filter) to all oscillatory components in its inputs. This is in contrast to the fact that each unit, in general, acts as a filter with non-uniform frequency response. Thus, the model can only facilitate correct identification of a source with a single frequency component, which is again unrealistic. To overcome this deficiency, an iterative post-processing algorithm that correctly identifies the physical source(s) is developed. An additional issue with the existing methods is that they lack a procedure to pre-screen non-oscillatory/noisy measurements which obscure the identification of oscillatory sources. In this regard, a pre-screening procedure is prescribed based on the notion of sparseness index to eliminate the noisy and non-oscillatory measurements from the data set used for analysis.

***Keywords*—** non-negative matrix factorization, PCA, source separation, plant-wide diagnosis

I. INTRODUCTION

The problem of source identification using multivariate spectral analysis has been widely studied in various domains such as chemometrics, speech processing, plant wide oscillation detection and astronomy [1], [2], [3], [4], [5]. Multivariate spectral analysis is also used in image analysis, classification and pattern recognition for understating the underlying phenomena [7], [5]. Several techniques have been used to identify of sources using the multivariate spectral data. Some of the widely used methods are Independent Component Analysis (ICA) and Nonnegative Matrix Factorization (NMF).

In the application of the above two methods to multivariate spectral analysis, the columns of the data matrix typically are the spectra of individual measurements. The spectra are obtained either by direct measurement [12], [7] or by Fourier transformation of the time domain measurements [9], [10], [8]. In this work the focus is only on the data obtained through latter means i.e. through Fourier transform of time domain data. However, when this spectral data is used for source identification in industrial processes, a linear mixing model is assumed. In the earlier works [9], [6], [8] the linear mixing model assumed for source identification is restrictive and is only valid if each unit of the process gives equal weighting (all-pass filter) to all oscillatory components in its inputs. Thus, the model can only facilitate correct identification of a source with a single frequency component, which is again unrealistic. Further, in both source identification methods (ICA and NMF), the number of sources should be known *apriori*. However, with the assumed mixing model the number of sources determination also poses difficulty due to the fact that the model gives equal weighting to all frequencies present in the data. The above issues are discussed in section II.

Principal Component Analysis (PCA) is a well known technique to detect the number of sources present in the multivariate data set. PCA exploits correlations among the multivariate process data to project the information in the original measurement space on to a reduce order space spanned by a set of orthogonal latent variables [2], [9], [10]. The dimension of the reduced space is an estimate of the number of sources. However, the estimate is correct only in the absence of measurement errors, non-linearities and noise [3]. It is to be noted that PCA is not a source identification technique.

Recently, a novel method to determine the number of sources for NMF based methods has been proposed by Tangirala et.al [8]. In their work the authors introduce the notion of a Pseudo Singular Value (PSV) value to determine the number of sources which is reviewed in III. In this work, PSV to determine the number of basis shapes.

Strength factor, a measure of the amount of basis shapes present in measurements which is defined by Tangirala et.al [8] and reviewed in III is used as a post-processing tool to overcome the restrictive nature of the assumed linear mixing model.

The main contribution of this work are *(i)* explaining the restrictive nature of the assumed linear mixing model for source identification in industrial proceses and *(ii)* development of a post-processing algorithm to overcome the restrictive nature of the assumed linear mixing model. Further, a pre-screening method is also used to eliminate the non-oscillatory/noisy measurements which obscure the

This work was not supported by any organization

A. K. Tangirala is with Faculty of Chemical Engineering, Indian Institute of Technology Madras, Chennai, India - 600 036 arunkt@iitm.ac.in

S. Babji is with the Department of Chemical Engineering, Indian Institute of Technology Madras, Chennai, India - 600 036 babggg1@gmail.com

identification of oscillatory sources.

The rest of the paper is organized as follows. Section II highlights the problems involved in determination of number of physical sources and its signature with the assumed linear mixing model. Section III describes the proposed methodology for determination and identification of physical sources present in the multivariate process data set. Simulation studies are presented in Section IV which demonstrate the practicality and utility of the proposed method. The article ends with a few concluding remarks in Section V.

II. Problem formulation: Restrictive nature of the existing linear mixing model

As discussed earlier, the existing linear mixing model for oscillatory source signature identification in processes gives equal weightage to all the frequencies present in the process output, thereby neglecting the frequency domain characteristics of the process. The assumed model is valid only if the process behaves as an all-pass filter which is mostly un-realistic. This is explained using the following simple example.

Consider measurements of 3 individual variables x_1, x_2 and x_3 from a process such that the time domain relationship is

$$x_3 = a_1 x_1 + a_2 x_2$$

Then their Fourier transforms are related through

$$X_3(\omega) = a_1 X_1(\omega) + a_2 X_2(\omega)$$

now, the relationship between the spectra of x_1, x_2 and x_3 is obtained as

$$\begin{aligned} |X_3(\omega)|^2 = \quad & a_{11}|X_1(\omega)|^2 + a_{22}|X_2(\omega)|^2 + \\ & 2a_1a_2|X_1(\omega)||X_2(\omega)|cos(\phi(\omega)) \end{aligned}$$

where $\phi(\omega)$ is the phase difference between $|X_1(\omega)|$ and $|X_2(\omega)|$ at frequency ω. The cross term $2a_1a_2|X_1(\omega)||X_2(\omega)|cos(\phi(\omega))$ is zero if one of the following two conditions is satisfied for all frequencies
(i) x_1 and x_2 have non overlapping spectra
(ii) x_1 and x_2 have overlapping spectra but $\phi(\omega) = \pi/2$ for all ω.

However, the above linear mixing model gives weightages a_1 and a_2 to the process data. This is in contrast to the fact that each unit, in general, acts as a filter with non-uniform frequency response. Therefore, the weights a_1 *and* a_2 *should depend on the frequency response of the process*. This is missing in the existing linear model and hence suited only for processes which acts as an all-pass filter. Further, the model can only facilitate correct identification of a source with a single frequency component, which is again unrealistic.

The assumption of equal weightage to all frequency components in the data poses difficulty in both determination of number sources and source signature identification. This is due to the fact that both frequency attenuation and amplification takes place in the process, thereby yielding wrong results with the usage of existing linear mixing model. These facts are explained using a simulation example given below:

Example:

A simulated system shown in Fig. 1 (a) is taken up to illustrate the aforementioned shortcomings. The process contains five LTI units (subsystems) with three closed-loops. The transfer functions of the units and controllers are given in Tables 1 and 2. Two sources of oscillations are considered, namely, a sticky valve (a 1-parameter model is used) and a tightly tuned controller. The former contains a spectral signature with multiple peaks while the latter source contains a signature with single peak (please see Fig. 1 (g)). Five variables are measured whose spectra are shown in Fig. 1 (c).

The spectral matrix is constituted in the traditional way and PCA is used to determine the number of sources. The number of non-zero singular values obtained from PCA with no noise in the data set is four $(0.49, 0.04, 0.02$ and $0.002)$ indicating that there are four sources. This highlights the fact that the determination of number of sources from the above linear mixing model leads to incorrect results. Further, the basis shapes obtained from NMF (with number of sources as 4) is shown in Fig. 1 (d). It is clear that the method not only fails to extract the correct source signatures but also fails to estimate the correct number of physical sources. The forthcoming section explains the proposed methodology with an example using NMF for source identification to overcome the above limitations in the linear mixing model.

III. Proposed methodology

As discussed in Section I, the shortcomings in the existing linear mixing model is overcome by the usage of a quantity known as strength factor [8]. The iterative post-processing algorithm is developed for source identification using NMF technique. This method which is used to overcome the above limitation of the model primarily relies on the assumption that the source is present amongst the measurements. In most cases this is easily satisfied since the control volumes are large enough to contain the source(s).

The steps involved in the proposed post-processing algorithm is given below:

1) Using the sparseness index, process variables that contain non-oscillatory components and noise elements are discarded. A threshold value of 0.8 is fixed for the sparseness index and all the basis shapes with lower sparseness index values are eliminated.

   Significance

   The sparseness index (definition adopted from paper by Patrik [11]) of a vector $'x'$ of dimensionality $'n'$ is given by

$$Sparseness(x) = \frac{\sqrt{n} - \left(\sum |x_j| / \sqrt{\sum |x_j|^2}\right)}{\sqrt{n} - 1}$$

   From the equation it is clear that the sparseness index decreases as the number of peaks and level of noise

increases in the variable and hence in this work, it is used as a pre-screening tool for source identification.

2) The spectra of the variables that are not eliminated using pre-screening technique is computed and a matrix containing the spectra of these variables is formed. NMF is run on this matrix with the absolute of initial values (Basis shapes B and weights W) obtained from singular value decomposition (SVD) of the spectral data matrix.
3) This step is concerned with the estimation of number of physical sources. Pseudo singular value defined in [8] is used to estimate the number of sources.

Significance

The notion of Pseudo Singular Value (PSV) was introduced in the context of NMF to determine the number of sources. Given a nonnegative matrix X, the PSV of j^{th} basis shape denoted by ρ_j and has been defined as,

$$\begin{aligned}\rho_j &= \frac{\left|\left|\sum_{k=1}^{j} C_k\right|\right|_2^2 - \left|\left|\sum_{k=1}^{j-1} C_k\right|\right|_2^2}{||X_T||_2^2}, \quad j = 2,..,L \\ &= \frac{||C_1||_2^2}{||X_T||_2^2}, \quad j = 1\end{aligned}$$

where L is the number of basis shapes and C_k, the k^{th} power component is given by [8]

$$C_k = \sum_{j=1}^{M} B_k W_{kj}$$

where B_k is the k^{th} basis shape and W_{kj} is the weight of the j^{th} variable corresponding to B_k obtained from NMF. X_T, the total power at each frequency ω_l is defined [8] as

$$X_T(\omega_l) = \sum_{j=1}^{M} |X_j(\omega_l)|^2$$

where $|X_j(\omega_l)|^2$ is the power spectrum of measurements at each frequency. The PSV signifies the amount of information captured by the basis shapes and ranges from 0 to 1 depending on the spectral data information that can be explained by the basis shapes.

4) This step is used to find the similarity between the measurements and the basis shapes using a measure known as Strength Factor (SF). SF has been defined in [8] and it is computed between the basis features of NMF and measurements.

Significance

Strength factor is a measure of the amount of basis shapes present in measurements and is given by the following equation

$$\begin{aligned}SF_{kj} &= \frac{\left|\left|\sum_{p=1}^{k} B_p W_{pj}\right|\right|_2^2 - \left|\left|\sum_{p=1}^{k-1} B_p W_{pj}\right|\right|_2^2}{||X_j||_2^2} \quad k = 2,..L \ \forall j \\ SF_{kj} &= \frac{||B_1 W_{1j}||_2^2}{||X_j||_2^2} \quad k = 1 \ \forall j\end{aligned}$$

where L is the number of basis shapes and X_j is the power spectrum of the j^{th} measurement [8]. The strength factor is a measure of the extent of the k^{th} *basis feature* in the j^{th} measurement.

5) Steps 2 to 4 are repeated with decreasing number of basis shapes untill each basis shape atleast corresponds to one single measurement.

Significance

As explained earlier, it is assumed that the source is present amongst the measurements. This gives rise to the fact that the *measurements corresponding to the sources* should completely explain the basis shapes while *the other measurements is a combination of the basis shapes*.

Demonstration of the proposed method using the example in Section II

(i) For the example described in Section II, the sparseness index of the third measurement is closer to zero indicating that this measurement is to be eliminated from the spectral data matrix, while all the other measurements have sparseness index values above the threshold. The plot of output variable x_3 shown in Fig. 1 (b) clearly reveals that the this output has no oscillations. Therefore, the spectral data matrix is now constituted of only four measurements namely x_1, x_2, x_4 and x_5.

*(ii)*The number of non-zero PSV obtained from the spectral data matrix in step 1 is three ($0.93, 0.05$ and 0.02), indicating three sources. However, this is not true since there are only two sources. Therefore, a measure known as strength factor (SF) is used in the next step to obtain the correct estimate of the number of sources and identify its signatures.

(iii) Strength Factor computed between the original measurements (without prescreening) and the basis shape feature is shown in Fig. 1 (e). It can be clearly seen that two of the basis shapes is not explained fully by the measurements. Therefore, an iterative procedure is needed to estimate the number of physical sources and identify its signatures.

(iv) Measurement three is eliminated by prescreening method and therefore, it is clear that the fourth basis shape is not completely explained by any of the measurement. Hence this particular basis shapes does not correspond to an individual source. Now, NMF algorithm is re-run with the number of basis shapes as two and the basis shapes are obtained. SF computed between the basis shape features and measurements is shown in Fig. 1 (f) which shows that each basis shape feature is explained fully by individual measurements. Therefore, the iterative procedure is complete and the basis

shapes obtained from NMF is shown in Fig. 1 (h). Notice that the basis shapes obtained correspond to the source signatures which is shown in 1 (g). Thus the number of sources are determined correctly along with its signatures.

IV. Simulation studies

In this section a simulated industrial process is taken up to demonstrate the potential of the proposed method.

Entech Data

The data set is from a simulated industrial process, courtesy of Entech control Inc.The simulated process shown in Fig. 2 (a). consists of a pulp manufacturing process, where the hardware and software pulps are mixed to give a stream of desired composition. The data set comprises of 1934 samples from 12 measurements associated with 12 control loops. It is desired to detect oscillations in the loops and to determine the physical sources present in the process.

Sparseness index computation reveals that measurment 2 can be eliminated from the data set since its value is below the threshold. A time plot of the data is shown in Fig. 2 (b). reveals that the measurement 2 is a noise element. Spectral data matrix is formed with the remaining 11 measurements and NMF algorithm is run initially with 9 sources. Three non-zero PSV values are obtained ($0.92, 0.06$ and 0.02) indicating three sources. Now to obtain the signature of these sources NMF is re-run with 3 sources.

SF values computed between the basis features and measurents.The SF plot shown in Fig. 2 (c). reveals that one of the basis shapes is not explained completely. Therefore, as explained in Section III, NMF algorithm is re-run with number of sources as 2. The strength factor computed between the basis shapes and measurements are shown in Fig. 2 (d). It clearly reveals the fact that there are only 2 sources in the process which is also confirmed from the knowledge of the process [8]. It has been stated in that valve stiction and poorly tuned controller are sources of oscillation in this process. The results obtained from the proposed method are in agreement with the results presented in earlier works. Further, the basis shapes obtained from the proposed method are shown inFig. 2 (e) along with the physical source signatures in Fig. 2 (f). It can be clearly seent that the physical source signatures are identified correctly using the proposed methodology.

V. Conclusions

This work brings into light the restrictive nature of the existing linear mixing model used for source identification from a multivariate process. It was shown that the existing linear model gives equal weightage to all the frequencies which can be used only if one of the following two conditions is satisified

(i) the process behaves as an all-pass filter.

(ii) there is only a single frequency component present in the input to the process.

A simulated example was used to demonstrate the restrive nature of the existing mixing model. A post-processing algorithm was developed to overcome the deficiency present in the model and to identify the source signatures. It was assumed that the sources are present amongst the measurements. In most cases this is easily satisfied since the control volumes are large enough to contain the source(s). An additional feature of the proposed method is the usage of sparseness index to remove non-oscillatory and noisy elements present in the multivariate data. It is to be noted that the elimination of variables using sparseness index is to avoid computation burden and the proposed method is insenstive to the sparseness index.

The proposed methodology is used to estimate the number of sources and its signatures. Diagnosis of the cause of oscillations by using the proposed method along with the plant topology is a subject of future work.

References

[1] S.M. Kanbur, D. Iono, N.R. Tanvir and M.A. Hendry. On the use of principal component analysis in analysing cepheid light curves. *Monthly Notices of the Royal Astronomical Society*, **329**(1):126–134, 2002.

[2] B.R. Bakshi. Multiscale PCA with application to multivariate statistical process monitoring. *AIChE*, **44**(7):1596–1610, 1998.

[3] D. Peter, W. Mitchell, and T. Lohnes. Maximum likelihood principal component analysis with correlated measurement errors: theoritical and practical considerations. *Chemometrics and Intelligent Laboratory Systems*, **45**(1):65–85, 1999.

[4] X. Li and X. Yao. Multiscale process monitoring in machining. *IEEE Transactions on Industrial Electronics*, **52**(3):924–925, 1998.

[5] D. Guillamet and J. Vitriz. A new iris recognition method using independent component analysis. *In Pattern Recognit. Lett.*,**24**, 2003.

[6] N.F. Thornhill and A. Horch. Advances and new directions in plant-wide disturbance detection and diagnosis. *Control Engineering Practice*, **15**(10):1196–1206, 2007.

[7] D.D. Lee and H.S. Seung. Learning the parts of object by nonnegative matrix factorization. *Nature*, **401**(3):788–791, 1999.

[8] A.K. Tangirala, J. Kanodia and S.L. Shah. Non negative matrix factorization for detection and diagnosis of plantwide oscillations. *Industrial Engineering and Chemistry Research*,**46**(3):801–817, 2007.

[9] N.F. Thornhill, S.L. Shah and B.Huang. Detection and diagnosis of unit wide oscillations. *Process Control and Instrumentation*, **26**, 2000.

[10] N.F. Thornhill, S.L. Shah, B.Huang and A.Vishnubhotla. Spectral principal component analysis of dynamic process data. *Control Engineering Practice*, **10**833–846, 2002.

[11] Patrik O. Hoyer. Nonnegative matrix factorization with sparseness constraints. *Journal of Machine Learning Research*, **5**14571469, 2004

[12] A.H. Rinen, J. Karhunen and E. Oja. Independent component analysis. *In Wiley, New York*, 2001.

$G_1(s)$	$G_2(s)$	$G_3(s)$	$G_4(s)$	$G_5(s)$
$\frac{1}{s^3+3s^2+3s+1}$	$\frac{3e^{-10s}}{10s+1}$	$\frac{3e^{-10s}}{10s+1}$	$\frac{4}{s+1}$	$\frac{1}{s+1}$

$C_1(s)$	$C_2(s)$	$C_3(s)$	V_2
$K_p = 3.7, K_I = 0.02$	$K_p = 0.1, K_I = 0.01$	$K_p = 0.1, K_I = 0.01$	*Stiction* $= 10\%$

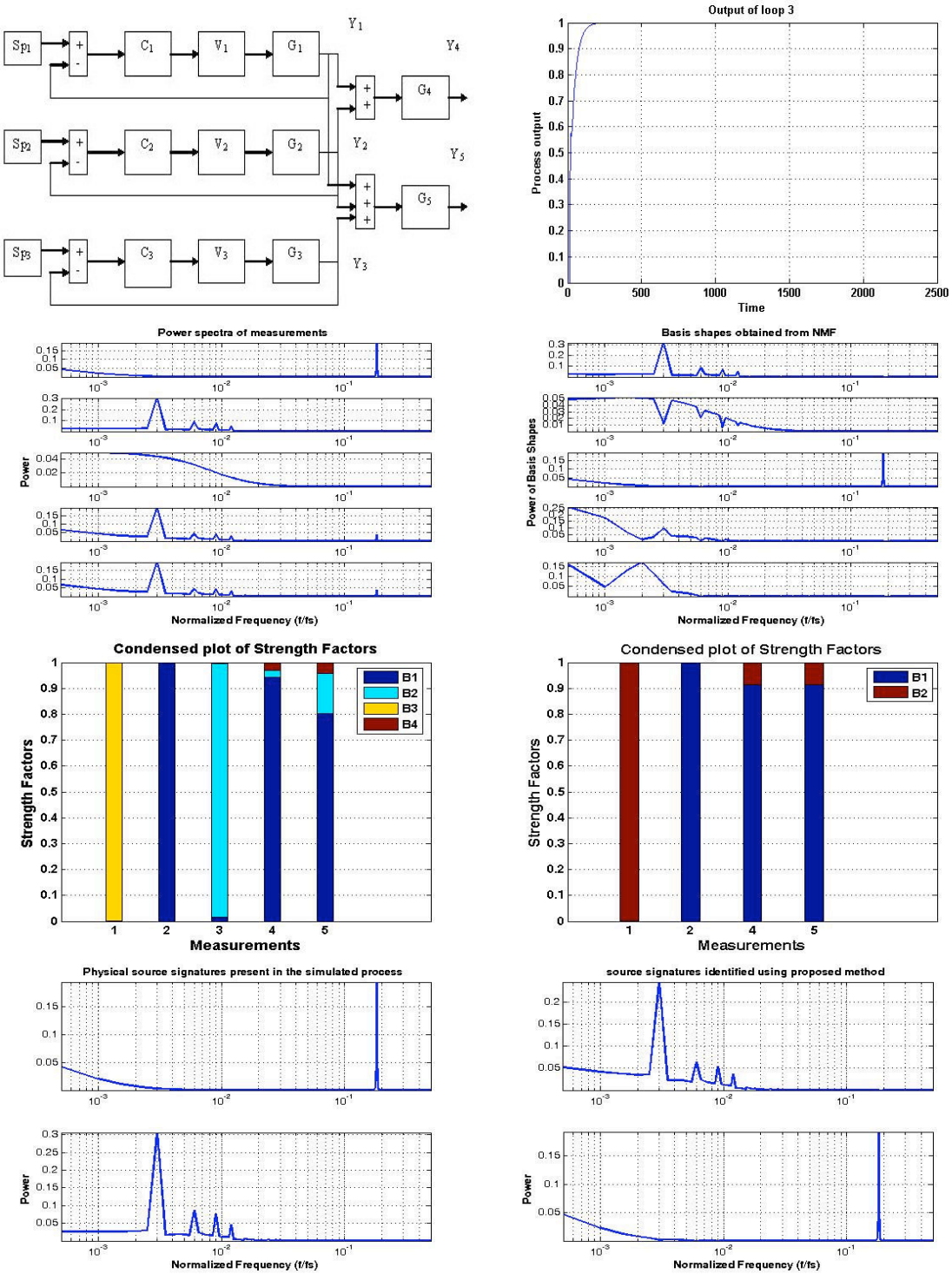

Fig. 1. (a) Simulated system (b) Process output from loop 3 in the linear system (c) Measurement spectra (d) Basis shapes from existing algorithm (f) Condensed plot of strength factors from traditional method (g) Condensed plot of strength factors from proposed method (h) Physical source signatures (e) Identified source signatures using proposed algorithm

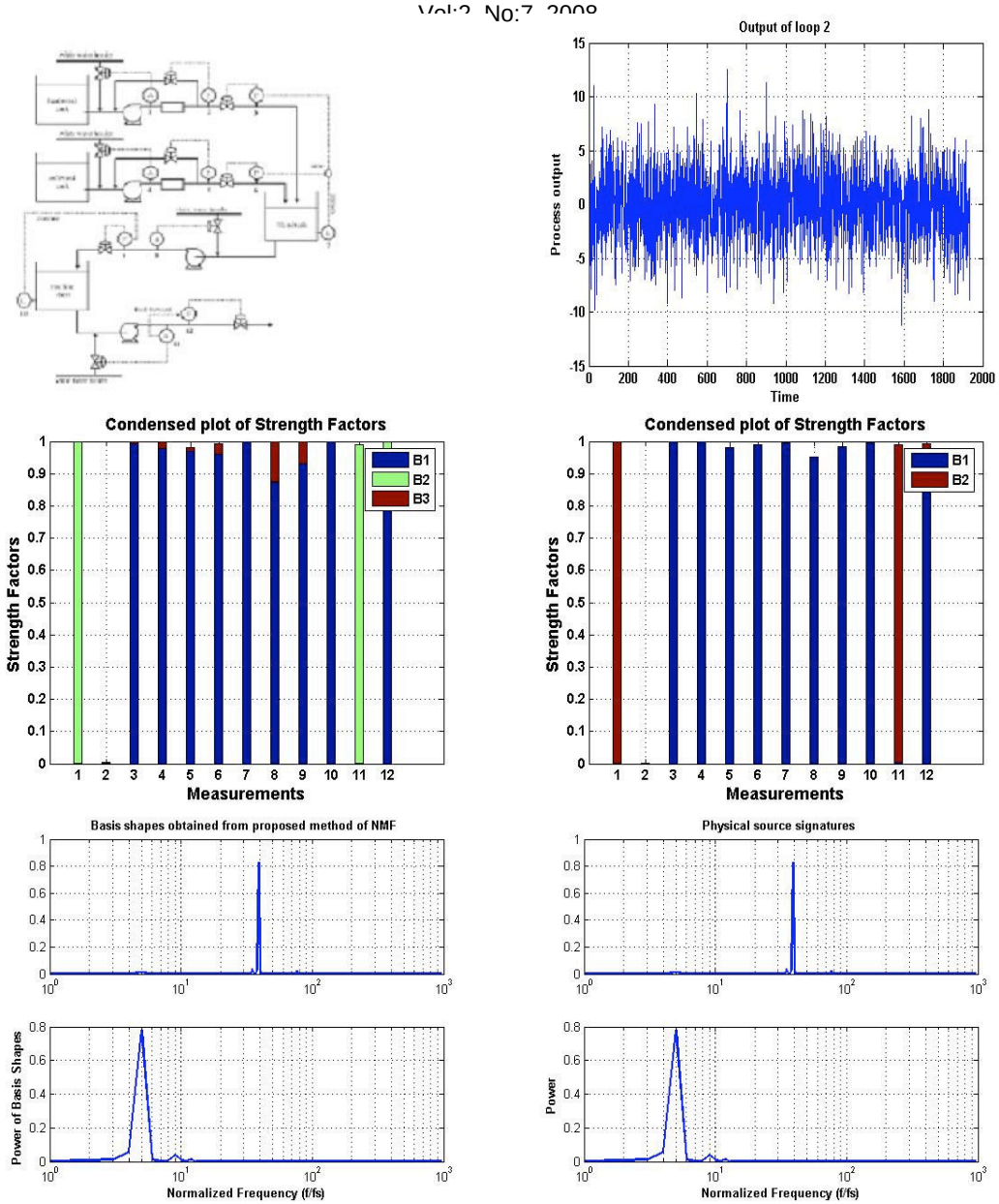

Fig. 2. (a) Schematic of Entech process (b) Process output from loop 2 (c) Strength factor obtained from traditional method (d) Strength factor obtained from proposed method (e) Identified source signatures using proposed algorithm (f) Physical source signatures