Proceedings of the 17th World Congress
The International Federation of Automatic Control
Seoul, Korea, July 6-11, 2008

Wavelet Based OE Model Identification with Random Missing Data

Lihui Geng, Tao Zhang, Deyun Xiao and Jingyan Song

Tsinghua National Laboratory for Information Science and Technology
Department of Automation, Tsinghua University
Beijing, China(e-mail: glh04@mails.tsinghua.edu.cn)

Abstract: Based on wavelet representation theory, this paper proposes a novel identification algorithm with random missing data under the condition that the identified dynamic process can be described as an output error (OE) model structure. This new algorithm mainly consists of two stages: one is the wavelet reconstruction, the other the prediction for missing data using the identified model. For the sake of its application, selection of the final iteration number and the adopted wavelet category is also considered. Finally, numerical simulations are given to verify the satisfactory effectiveness of the proposed algorithm.

Keywords: Recursive identification; Time series modelling.

1. INTRODUCTION

The phenomenon of random missing data is widespread in the process system, which may result from malfunction of the corresponding actuators or/and measurement instruments. So how to identify the required process control model in the presence of missing data is a paramount task.

Up to now, a large number of identification algorithms have been proposed by many researchers, and Little [1987] has given a good summary of statistics methods with missing data. In the past few years, identification with missing data has gained much attention in control field where a good few identification methods having promising applications are proposed, see Isaksson [1993], Pintelon et al. [2000], Albertos et al. [1999], and Sanchis et al. [2002]. In Isaksson [1993], expectation maximum (EM) algorithm is derived to cope with ARX model identification with both input and output data missing. As a matter of fact, it is a recursive maximum likelihood estimation (MLE) but with approximate 1/4 computational load compared with that of MLE. Pintelon et al. [2000] propose an errors-in-variables (EIV) model identification approach in frequency domain, it can be applied to various model structures, and is suitable to systems persistently excited by all forms of signals. Besides the above merits, it can also be used for unstable systems. The main drawback of this method lies in its time-consuming computation when the number of missing data is large, since it regards the missing data as parameters to be identified. Albertos et al. [1999] and Sanchis et al. [2002] study a simple identification algorithm, which is based on the prediction for missing data using the identified model.

Wavelet analysis technique (Daubechies [1992]) is prevalent in the field of image processing, information compression and denoising. It has a fine feature of time-frequency localization embodied by all sorts of mother wavelet and scaling functions which can produce the orthonormal basis functions for signals in space $L^2(R)$ (Mallat [1989]). In this paper, the wavelet reconstruction technique is employed to filter the data so that the statistical property of signals are greatly improved, which is beneficial to the subsequent identification and prediction for missing data.

The layout of this paper is as follows. Section 2 explains why the mathematical model formulation focuses on OE model structure while the novel identification algorithm is proposed in section 3. Selection of the final iteration number and the adopted wavelet category is analyzed in section 4. We resort to section 5 to verify the effectiveness of the proposed algorithm by some numerical simulations. Finally, a conclusion is arrived at in section 6.

2. PROBLEM STATEMENT

According to the research results in Palavajjhala et al. [1996], wavelet reconstruction is equivalent to a filtering to $y(k)$, $u(k)$ and $e(k)$ respectively. After the filtering is performed at some specific level, the statistical properties of the filtered signals will become more suitable for process identification due to the further improved input or output S/N ratio.

In order to make full use of the wavelet reconstruction technique to obtain satisfactory prediction results for the missing data, the OE model structure is considered, which has the following parameterized form (Ljung [2002]):

$$y(k) = \frac{B(q^{-1})}{F(q^{-1})}u(k) + e(k) \tag{1}$$

where q stands for the forward shift operator; k is short for kT_s, T_s being the sampling period of the system; $B(q^{-1}) = b_1q^{-1} + b_2q^{-2} + \cdots + b_{n_b}q^{-n_b}$ and $F(q^{-1}) = 1 + f_1q^{-1} + \cdots + f_{n_f}q^{-n_f}$; $y(k)$ represents the measured output; $u(k)$ is the reference input and $e(k)$ a measurement white

noise.

By means of carefully choosing the wavelet decomposition level, the variance of filtered $e(k)$ will be reduced more greatly than that of the filtered undisturbed signal $s(k) = B(q^{-1})/F(q^{-1})u(k)$. Therefore, by this way, the maximal output S/N ratio can be gained to improve the accuracy of identified model. Besides, the OE model has a fine feature of global and local identifiability and consistent convergence under the quite mild conditions. As a result, the identified OE model will ensure a high precision of the subsequent prediction for missing data. So in this paper, the OE model frame is adopted to formulate the process dynamic model.

3. IDENTIFICATION ALGORITHM

When the data are lost randomly, the wavelet reconstruction on its own is determined not to predict the missing data reasonably. This is due to the reason that it can not provide any further information about the identified process. Therefore, this reconstruction technique should be combined with other prediction methods.

The simplest way to predict the random missing data may be to make full use of the identified model. So the prediction equation can be written as:

$$\hat{x}(t) = \hat{f}(Y^{t-1}, U^{t-1}) \tag{2}$$

where $\hat{x}(t)$ is the prediction for input or output missing data $x(t)$, and Y^{t-1}, U^{t-1} represent the output and input data not greater than instant $t-1$ respectively. $\hat{f}(\cdot,\cdot)$ is the identified model.

However, as was stated in Albertos et al. [1999], the main drawback of this means is the lack of robustness to noise and external disturbances, which makes this simple method hard to use. In this section, the wavelet reconstruction technique is combined with this model based prediction strategy to improve the prediction ability for the missing data.

The novel identification algorithm coping with missing output data includes the following analysis stages:

3.1 Initial interpolation for missing data

For the purpose of wavelet reconstruction at the later stage, the missing data should be at first filled in by means of interpolation.

3.2 Reconstruction of whole data

To realize data reconstruction with the specified resolution, a level q should be properly chosen to approximate the whole data including sampled data as well as the interpolated or predicted ones for missing data. Therefore we have the following decomposition expression:

$$\hat{y}(k) = \sum_{j=1}^{l_q} \phi_{q,j}(t_k) a_{qj} + v(k) \tag{3}$$

where $\phi_{q,j}(t_k) = 2^{-q/2} b_0^{-1/2} \psi(2^{-q} b_0^{-1}(t_k - j2^q b_0))$ is the wavelet basis function at level q where $\psi(\cdot)$ is any mother wavelet; j denotes a nonnegative integer; $b_0 > 0$ is an arbitrary chosen constant; $v(k)$ is the approximation error by wavelet representation; $\hat{y}(k)$ represents the output data sampled or predicted at instant t_k.

So the reconstructed output data using wavelet are $\hat{\hat{y}}(k) = \sum_{j=1}^{l_q} \phi_{q,j}(t_k)\hat{a}_{qj}$. Similarly, the input data reconstructed by wavelet have the same reconstructing process as $\hat{\hat{y}}(k)$, which are $\hat{\hat{u}}(k) = \sum_{j=1}^{l_q} \phi_{q,j}(t_k)\hat{b}_{qj}$.

3.3 Identification of model parameters

According to the filtered data by wavelet, the identified model takes an OE model structure:

$$\hat{\hat{y}}(k) = \frac{B(q^{-1})}{F(q^{-1})}\hat{\hat{u}}(k) + e(k) \tag{4}$$

where $e(k)$ is a white noise while $\hat{\hat{y}}(k)$ and $\hat{\hat{u}}(k)$ are the reconstructed data by wavelet in subsection 3.2.

3.4 Prediction for random missing data

In terms of (4), the best prediction for missing output data will be:

$$\check{y}(k) = \frac{\hat{B}(q^{-1})}{\hat{F}(q^{-1})}\hat{\hat{u}}(k) \tag{5}$$

where $\hat{B}(q^{-1})$ and $\hat{F}(q^{-1})$ are respectively the numerator polynomial and the denominator polynomial of the identified process model in subsection 3.3.

To improve the precision of the predicted output data and parameter identification results afterwards, the above subsection 3.2, 3.3, and 3.4 can be iterated several times.

As far as the implementation of the above algorithm is concerned, it can be decomposed as the following steps:

Step 1: Linear interpolation for the missing data

Step 2: Reconstruct the whole data with the wavelet basis functions at level q by means of least square fitting:

$$\hat{A}_q = [\Gamma^T\Gamma]^{-1}\Gamma^T\hat{Y}^N \tag{6}$$

where $\hat{A}_q = [\hat{a}_{q1}, \hat{a}_{q2}, \cdots, \hat{a}_{ql_q}]^T$,$\Gamma = [\Gamma_{q1}, \Gamma_{q2}, \cdots, \Gamma_{ql_q}]$ $\Gamma_{qs} = [\phi_{q,s}(t_1), \phi_{q,s}(t_2), \cdots, \phi_{q,s}(t_N)]^T (s = 1, \cdots, l_q)$ and $\hat{Y}^N = [\hat{y}(1), \hat{y}(2), \cdots, \hat{y}(N)]^T$. Therefore the reconstructed output data can be obtained by:

$$\hat{\hat{y}}(k) = e_k\Gamma\hat{A}_q \tag{7}$$

where e_k is a unit row vector whose kth element is 1 and the other elements are 0s. Likely, $\hat{\hat{u}}(k)$ is reconstructed in the same way as $\hat{\hat{y}}(k)$.

Step 3: Recursive identification of the parameterized process model (c.f. Ljung et al. [1983]):

Initialize $\hat{\theta}(0)$, $R(0)$ and then iterate the following equations:

$$\check{y}(k) = h^T(k)\hat{\theta}(k-1) \tag{8}$$

$$\tilde{y}(k) = \hat{y}(k) - \check{y}(k) \tag{9}$$

$$R(k) = R(k-1) + \frac{1}{k}[\psi(k)\psi^T(k) - R(k-1)] \tag{10}$$

$$K(k) = \frac{1}{k}R^{-1}(k)\psi(k) \tag{11}$$

$$\hat{\theta}(k) = \hat{\theta}(k-1) + K(k)\tilde{y}(k) \tag{12}$$

where $\hat{\theta}(k) = [\hat{b}_1(k), \cdots, \hat{b}_{n_b}(k), \hat{f}_1(k), \cdots, \hat{f}_{n_f}(k)]^T$, $h(k) = [\hat{\hat{u}}(k-1), \cdots, \hat{\hat{u}}(k-n_b), -u^*(k-1), \cdots, -u^*(k-n_f)]^T$, $u^*(k) = h^T(k)\hat{\theta}(k)$, $\psi(k) = [u_f(k-1), \cdots, u_f(k-n_b), -u_f^*(k-1), \cdots, -u_f^*(k-n_f)]^T$, $u_f(k) = \hat{\hat{u}}(k) - \hat{f}_1(k)u_f(k-1) - \cdots - \hat{f}_{n_f}(k)u_f(k-n_f)$ and $u_f^*(k) = u^*(k) - \hat{f}_1(k)u_f^*(k-1) - \cdots - \hat{f}_{n_f}(k)u_f^*(k-n_f)$.

The final precision of estimated parameters in this step is determined by the data length N used for model identification.

Step 4: The missing output data are therefore predicted according to the final identified parameters $\hat{\theta}(N)$:

$$\check{y}(k) = h^T(k)\hat{\theta}(N) \tag{13}$$

Step 5: Substitute $\hat{y}(k)$ in step 2 with $\check{y}(k)$ in step 4 and iterate step 2, 3, 4, and 5 several times. It should be noted that the different wavelet basis functions may be employed in step 2 of each iteration.

Remark: Wavelet reconstruction with basis functions in this algorithm acts as a low pass filter. However, the adopted wavelet reconstruction technique outperforms the low pass filter since no extra gain and phase are added to the identified process model. By elaborately selecting the specific wavelet, the aforementioned algorithm has a good property of robustness to noise and shows an effective prediction ability, which will be explained by the simulations and comparisons in section 5.

4. SELECTION OF ITERATION STEP AND WAVELET CATEGORY

For the purpose of its application, the algorithm presented in section 3 should be further studied on how to determine the proper iteration number and wavelet category in each iteration. Therefore, in this section, these two aspects will be discussed accordingly.

4.1 Determination of the appropriate wavelet category in each iteration

As was stated in Palavajjhala et al. [1996], the input or output S/N ratio had been used as a performance criterion to design the prefilter using wavelet transform in order to improve the accuracy of identified model. The research results there show that the S/N ratio can be a good indicator for selection of wavelet category. Since the newly proposed algorithm assumes that the output data are missing, the S/N ratio of predicted output data is much important than that of input data. Therefore, it is naturally chosen as a criterion to determine the wavelet category employed in the iteration.

Since the wavelet reconstruction technique is employed to filter the whole data in step 2, the output S/N ratio of these data should be computed before the wavelet analysis. Nevertheless, the predicted data in each iteration are related to many factors, which make it difficult to obtain the ratio. Considering its computation, it is therefore assumed that the output S/N ratio does not change greatly before and after the prediction for missing data. Under this assumption, the ratio can be calculated conveniently by the recursive identification results in step 3.

The approximate output S/N ratio calculated by means of the above method can be the feasible criterion for selecting the wavelet category for data reconstruction, which will be demonstrated in the subsequent simulation study.

4.2 Determination of the final iteration number

Compared with the selection of wavelet category, the choice for final iteration number of this algorithm is much easier. Generally, the most often used criterion for stopping the iteration in model identification is relevant to the accuracy of identified parameters. Hence

$$\|\hat{\theta}_i(N) - \hat{\theta}_{i-1}(N)\|_2 < \epsilon \tag{14}$$

can be taken as a feasible stop criterion, where ϵ is a small positive value specified by the user and i the ith iteration of this algorithm.

5. SIMULATION STUDY

In this section, simulations will be given for verification of the presented algorithm and the specified selection in section 4.

In our simulation, a two order time-invariant attitude model of small satellites, linearized at a specific operating point, is considered:

$$G(z) = \frac{0.03259z + 0.031}{z^2 - 1.85z + 0.8607}. \tag{15}$$

The exciting reference input is a four-period PRBS with total data length 248, whose amplitude is 10. In these simulations, the input and output data are both sampled with a period T_s, being 0.125s, which is a half of the clock time T_c of the PRBS. For the purpose of identification, the first period of PRBS is discarded and the others are left.

In the course of simulation, the following two main points should be taken into account:

5.1 Ill effect caused by the initially chosen wavelet category and initial linear interpolation for missing data

Since wavelet reconstruction (step 2) is prior to model parameter identification (step 3) in the initial iteration, the arbitrarily selected wavelet category will in no doubt have ill effect on the subsequent model identification and prediction for missing data (step 4). Moreover the much more important influence is the initial estimation errors caused by the guessed linear interpolation for missing data in step 1. Consequently, a transient process will be shown in the predicted data, which is unsuitable for identification of time invariant model. The predicted data in this transient process should not be used for identification and therefore should be excluded according to the priori settling time of the identified process.

5.2 Selection of the level for wavelet reconstruction

In terms of the sampling period T_s, the predicted output signal is concentrated on the frequency interval [0, ω_s], where $\omega_s = \pi/T_s = 25.133 rad/s$. According to the priori crossover frequency ω_c of the dynamic process, the cutoff frequency of the low pass filter realized by wavelet reconstruction should be greater than ω_c. Therefore the resolution of wavelet reconstruction is selected as $q = 1$, which corresponds to a filtering frequency band between 0 and $\omega_s/2 = 12.566 rad/s$.

Before the simulations, some notations are defined for convenience: WBOEP stands for the wavelet based OE model prediction while OEP for the OE model prediction. Moreover, the prediction methods and iteration procedures in both algorithms are mainly the same, except that whether there is the participation of the wavelet reconstruction or not. In all these simulations, the Daubechies (including haar wavelet) wavelets are adopted.

Simulation1: In this simulation, suppose that there are approximately 80% missing in output data and that the undisturbed output data are corrupted by a white noise $e(k)$ with zero mean and variance 0.1.

After the implementation of the algorithm, the changes of prediction errors with the increasing iteration number are shown in Fig. 1. The data before the time instant 18.625s are predicted errors while the remaining data belongs to the wavelet reconstruction errors. As can be seen from this graph, the prediction errors undergo a process from the large error at the beginning of the iteration to the small one at the end of iteration. This changing process of decreasing prediction errors suggests the feasibility of the means of computing the output S/N ratio mentioned in section 4. Also, the large prediction errors influenced by the initial linear interpolation and the initially chosen wavelet category are illustrated in this figure. Consequently, the data ranging from 0 to 4.5s are thrown away to improve the accuracy of model identification and prediction afterwards. The predicted data using these two different algorithms are compared in Fig. 2 and Fig. 3. Fig. 2 makes a comparison of the whole data including the predicted data as well as the sampled data while Fig. 3

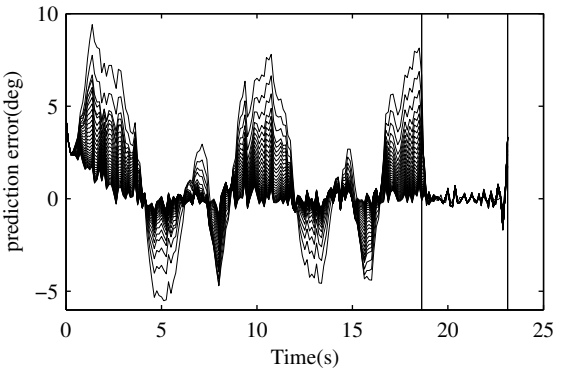

Fig. 1. Evolution of prediction errors with iteration number

gives a locally magnified graph of the predicted data in Fig. 2. The same effect due to the initial linear interpolation is also illustrated in Fig. 2. It is shown from Fig. 3 that the WBOEP data are much more close to the real data than that of OEP data. With these predicted data, the process models are identified respectively. Fig. 4 and Fig. 6 separately show the amplitude-frequency and phase-frequency comparisons between the identified models by employing different algorithms. Fig. 5 and Fig. 7 respectively are the locally magnified graphs of Fig. 4 and Fig6. Besides the performance comparison in frequency domain, Fig. 8 gives the step response comparison of these identified models. The changes of selected wavelet categories in iterations

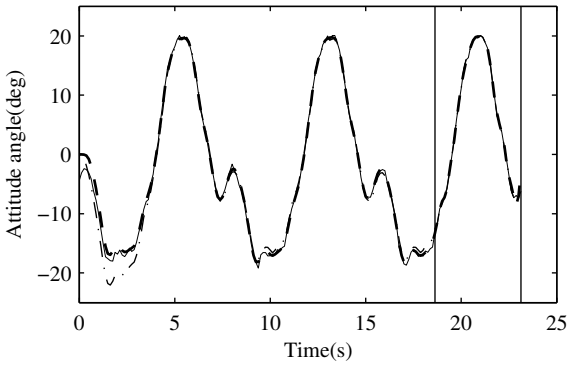

Fig. 2. Comparison between WBOEP data (dashed line), OEP data (dashdot line) and real sampled data (solid line)

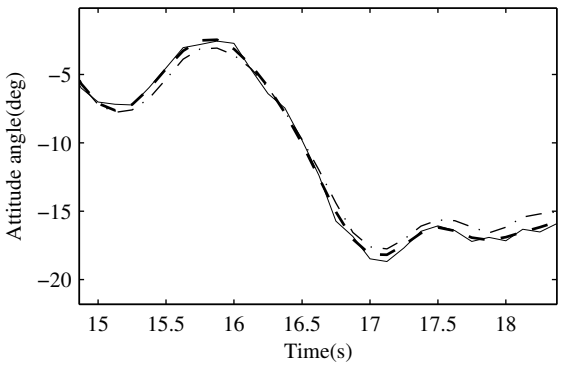

Fig. 3. A locally magnified graph of Fig. 2

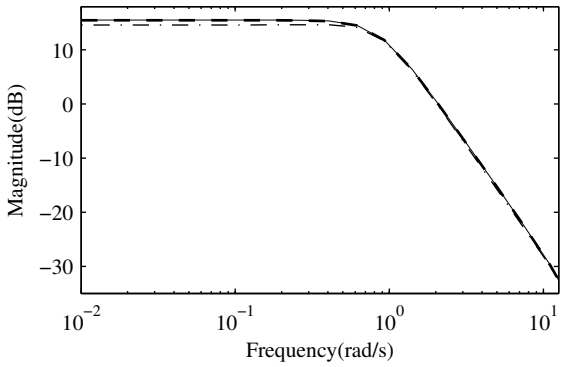

Fig. 4. Amplitude-frequency comparison between identified models with WBOEP data (dashed line), with OEP data (dashdot line) and real attitude model (solid line)

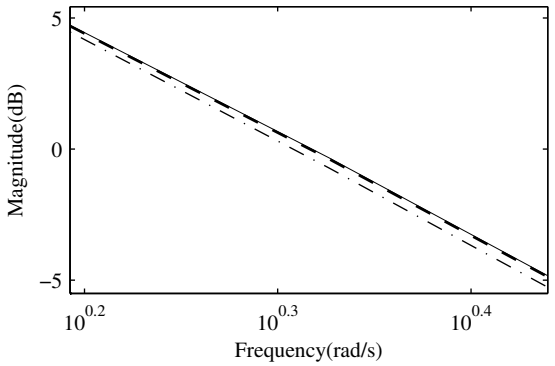

Fig. 5. A locally magnified graph of Fig. 4

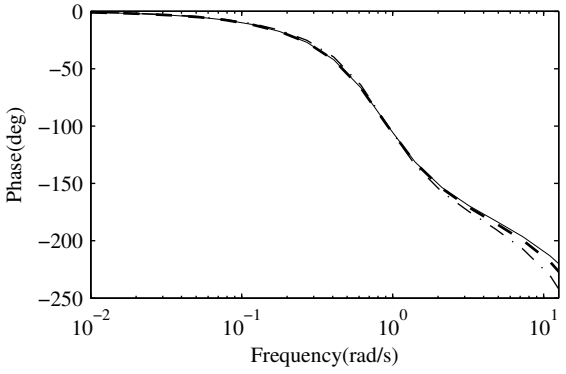

Fig. 6. Phase-frequency comparison between identified models with WBOEP data (dashed line), with OEP data (dashdot line) and real attitude model (solid line)

are shown in Fig. 9. As can be seen from this figure, from iteration number 1 to 3, db3 wavelet is employed while from iteration number 4 to 5, db9 wavelet is adopted. db8 wavelet is utilized in the remaining iterations until the termination of the algorithm.

Conclusion 1: The proposed algorithm is superior to the one employing OE model prediction without wavelet reconstruction in the absence of approximate 80% data.

Simulation 2: In order to see if this newly presented algorithm is preferable to its counterpart all the time,

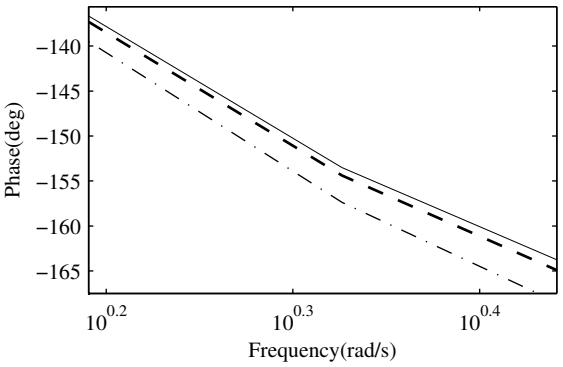

Fig. 7. A locally magnified graph of Fig. 6

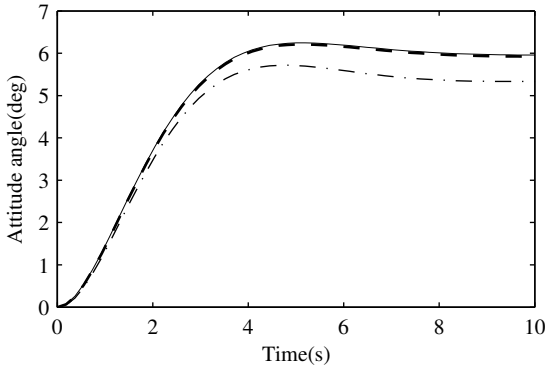

Fig. 8. Step response comparison between identified models with WBOEP data (dashed line), with OEP data (dashdot line) and real attitude model (solid line)

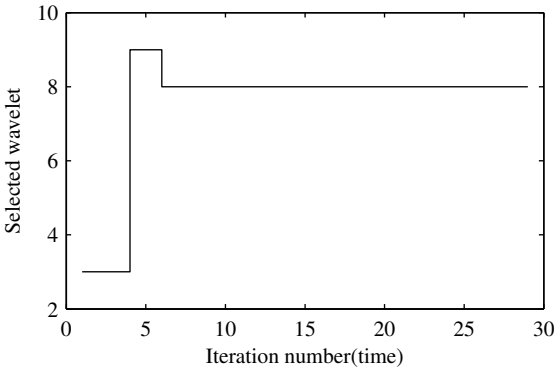

Fig. 9. Selected wavelet categories in iterations

the amount of output missing data is varied. By sufficient simulations and comparisons, the following conclusions can arrived at.

Conclusion 2: When a small number of output data are lost, the final iteration number will decline providing ϵ remains the same. Meanwhile, these two algorithms give the corresponding identified models with approximate accuracy (Fig. 10). However, conversely, when there are a large number of missing data, the proposed algorithm will outperform its counterpart by yielding an identified model with higher accuracy (Fig. 11).

Simulation 3: By varying the variance of output noise, the simulations are made to see whether the novel algorithm is robust in the presence of 80% missing data. After these simulations, the following results can be obtained.

Conclusion 3: By elaborately selecting wavelet basis functions, the good prediction and identification results will be obtained, which suggest this proposed algorithm has the robustness to the noise with a small variance, ranging from 0 to 1 (Fig. 12 and Fig. 13).

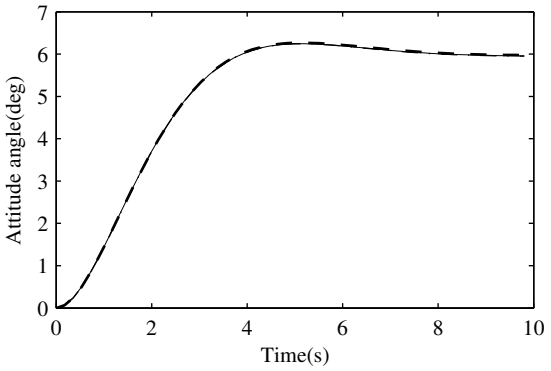

Fig. 10. Step response comparison of identified models in the presence of 25% missing data (with 2 iterations)

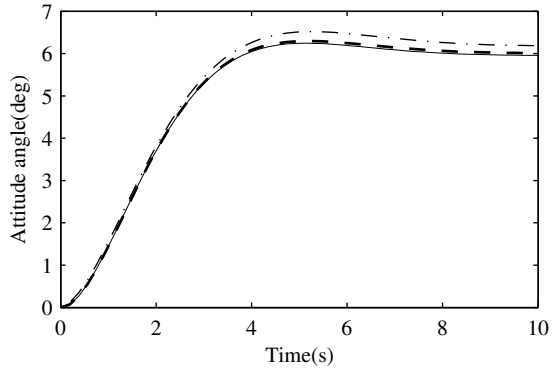

Fig. 11. Step response comparison of identified models in the presence of 75% missing data (with 13 iterations)

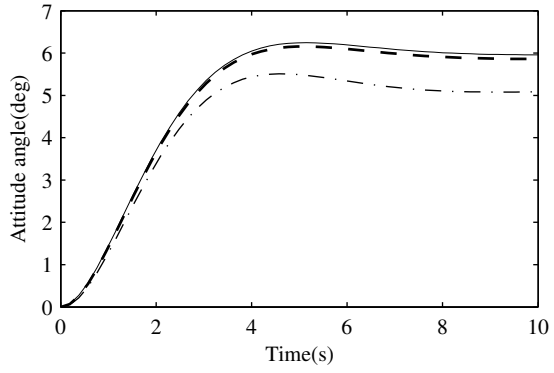

Fig. 12. Step response comparison of identified models with output noise variance 0.05

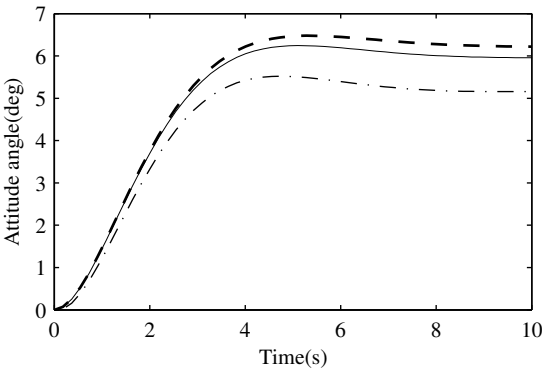

Fig. 13. Step response comparison of identified models with output noise variance 1

6. CONCLUSIONS

In this paper, a novel OE model identification algorithm is proposed to cope with the random missing data. The wavelet reconstruction technique is employed to improve the prediction ability for the missing data. Moreover, for the implementation of this algorithm, the selection of wavelet category in each iteration, the final iteration number, and other concerned parameters is also involved. Through plenty of simulations and comparisons, the conclusion can be therefore drawn that our proposal is superior to the algorithm without wavelet reconstruction when a certain number of data are missing.

REFERENCES

R.J.A. Little. and D.B. Rubin. Statistical Analysis with Missing Data. Wiley, New York, 1st edition, 1987.

A. J. Isaksson. Identification of ARX models subject to missing data. *IEEE Transactions on Automatic Control*, 38:813–819, 1993.

R. Pintelon and J. Schoukens. Frequency domain system identification with missing data. *IEEE Transactions on Automatic Control*, 45:364–369, 2000.

P. Albertos, R. Sanchis and A. Sala. Output prediction under scarce data operation: control applications. *Automatica*, 35:1671–1681, 1999.

R. Sanchis and P. Albertos. Recursive identification under scarce measurements-convergence analysis. *Automatica*, 38:535–544, 2002.

I. Daubechies. Ten Lectures on Wavelets. CBMS-NSF Regional Series in Applied Mathematics, SIAM, 1992.

S. G. Mallat. A theory for multiresolution signal decomposition: the wavelet representation. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 11:674–693, 1989.

L. Ljung. System Identification: Theory for the User. Tsinghua University Press, Beijing, 2nd edition, 2002.

S. Palavajjhala, R. L. Motard and B. Joseph. Process identification using discrete wavelet transforms: design of prefilters. *J. AIChE*, 42:777–790, 1996.

L. Ljung and T. Söderström. Theory and Practice of Recursive Identification. MIT press, Cambrige, Massachusetts, 1983.