21th IMEKO TC4 International Symposium on Understanding
the World through Electrical and Electronic Measurement,
and 19th International Workshop on ADC Modelling and Testing
Budapest, Hungary, September 7-9, 2016

FFT-based Identification of Data Loss Models

László Sujbert and György Orosz

Department of Measurement and Information Systems, Budapest University of Technology and Economics, Budapest, Magyar Tudósok krt. 2, Hungary, {sujbert,orosz}@mit.bme.hu

***Abstract* – Recently measurement data loss has been of greater interest, due to the spread of sensor networks and the idea of Internet of things. A procedure is proposed that is able to identify the most frequently employed data loss models. It is assumed that the communication protocol provides information about data loss, i.e. the so-called data availability indicator function is known. The power spectral density (PSD) of the indicator function is representative for the model, and can be used for identification. Spectral estimation is carried out by Fast Fourier Transform (FFT) based techniques. The paper introduces the identification procedure for random independent, random block-based and a Markov model-based data loss patterns. The efficiency of the proposed method is demonstrated by simulation and measurement results.**

***Keywords*—** data loss; measurement; FFT; PSD; system identification

I. INTRODUCTION

Nowadays measurement data transfer is frequently carried out in sensor networks or on the Internet. In this case data can be corrupted or the transmission medium can be partially damaged, etc. [1],[2]. The presence of such systems motivated the investigation of data loss phenomena from signal processing point of view. Our recently published paper [3] discussed one of such problems: the handling of data loss in the case of spectrum estimation. In spite of the existing methods concentrating to the spectrum estimation, this paper dealt with the characterization of distortion caused by missing data.

Spectrum estimation based on time records with irregular sampling has been used for a long time. Records with missing data can be treated as a special irregular sampling, substituting the signal samples with zeros where the data are lost. Theoretically, such records can be synthesized by the multiplication of the original signal (without data loss) and the so-called data availability indicator function. The latter equals unity everywhere, with the exception of the time instants where data are missing, where it is zero. The spectral estimator of the damaged record is the discrete convolution of the original spectrum and the spectrum of the data availability indicator function.

Thus the introduction of the distorted spectra involved the calculation of the spectra of the data availability indicator functions. Three data loss models have been investigated: random independent, random independent block-based, and Markov model-based data loss. All their spectra have been determined, and quantitative connection between the data loss model parameters and the spectral parameters have been calculated.

In this paper we propose the inverse procedure: the data loss model can be identified by the Fourier transform of the data availability indicator function. First the PSD of the indicator function is to be calculated, then a parametric system identification method is to be used to get the spectral parameters. As the spectral shape is quite simple, this step is not critical. The last step is the calculation of the data loss parameters by the already known relations.

In section II the data loss models and their spectra are introduced in detail. Section III deals with the identification procedure itself, while section IV presents simulation and measurement results that confirm the procedure in practice, as well.

II. SPECTRUM ESTIMATION IN THE CASE OF DATA LOSS

A. Power Spectrum Estimation

The Fourier transform of a sampled signal $x(t_n)$ can be estimated by a finite set of samples [4]. The signal $x(t)$ is usually equidistantly sampled, and the spectrum is calculated by the Discrete Fourier Transform (DFT):

$$X(f_k) = \sum_{n=0}^{N-1} x_n e^{-j\frac{2\pi}{N}nk}, \;\; n,k = 0 \ldots N-1, \qquad (1)$$

where $f_k = k/N$ and $x_n = x(t_n)$. The DFT of a signal is usually calculated by the computationally efficient Fast Fourier Transform (FFT). The transformed vector $X(f_k)$ is generally complex valued, and the spectral content of the signal is expressed by the real valued Power Spectral Density (PSD) function:

$$S(f_k) = \frac{1}{N}|X(f_k)|^2. \qquad (2)$$

In order to reduce the variance of the PSD, a long series of samples is recorded, and many consecutive blocks of N samples are transformed, and the estimator is obtained by averaging the individual PSDs. The mean of the individual estimates can be calculated either by linear or exponential averaging.

B. Formulation of Data Loss

In order to model the data loss, a so-called data availability indicator function, K_n, is introduced [5]:

$$K_n = \begin{cases} 1, & \text{if the sample is processed at } n \\ 0, & \text{if the sample is lost at } n \end{cases}, \quad (3)$$

Samples which are not lost will be termed as processed or available samples. The data loss rate can be defined with K_n as:

$$\gamma = P\text{rob}\left\{K_n = 0\right\}, \quad (4)$$

where $P\text{rob}\left\{\cdot\right\}$ stands for the probability operator. The probability that a sample is available is $\mu = 1 - \gamma$.

C. Spectrum Estimation with Missing Data

Using the indicator function, K_n, (1) can be rewritten for the case of data loss:

$$\begin{aligned} \hat{X}(f_k) &= \text{DFT}(x_n K_n) = \sum_{n=0}^{N-1} x_n K_n e^{-\jmath \frac{2\pi}{N} nk}, \\ n, k &= 0 \ldots N-1, \end{aligned} \quad (5)$$

This formula means that by incorporating K_n into the usual form of DFT, missing samples are practically substituted with zeros. Equation (5) can also be evaluated via FFT.

The spectrum of the signal containing missing samples is obtained as the convolution of the spectrum of the lossless signal and the spectrum of the data loss indicator function. Now only the latter is interesting. Let $X_K(f_k)$ denote the Fourier transform of the data loss indicator function:

$$X_K(f_k) = \text{DFT}(K_n), \quad (6)$$

Thus the PSD of the data loss indicator function is:

$$S_K(f_k) = \frac{1}{N} |X_K(f_k)|^2. \quad (7)$$

The variance of $S_K(f_k)$ can also be reduced by averaging.

D. Data Loss Models and their Spectra

In paper [3], three data loss models have been investigated:

1. random independent data loss,
2. random block-based data loss,
3. Markov model-based data loss.

The random data loss is one of the most essential data loss models, it is often used because of its simplicity [6]. Block-based data loss models are often used, e.g., when several measurement results are transmitted over packet-based communication systems. Markov model has been proven to be useful, e.g., in the description of data loss pattern in real-time data transmission over Internet [7].

D..1 Random Independent Data Loss

Random independent data loss can be defined as follows:

$$\begin{aligned} K_n &= 1, \quad \text{with probability } \mu = 1 - \gamma \\ K_n &= 0, \quad \text{with probability } \gamma \end{aligned} \quad \text{for} \quad \forall n. \quad (8)$$

The definition means that each sample is lost with probability γ, and data losses at different time instants are independent of each other. The PSD of the data loss pattern is [3]:

$$S_K(f_k) = \frac{\mu(1-\mu)}{N} + \mu^2 \delta(f_k), \quad (9)$$

where $\delta(f)$ stands for the Dirac-delta function. The PSD is white, which is represented by the first term, while the second term represents the power of the mean value μ of the data loss pattern.

D..2 Random Block-based Data Loss

To define the random block-based data loss, the indicator function is given as:

$$\begin{aligned} \{K_{kM} \ldots K_{(k+1)M-1}\} &= 1, \quad \text{with probability } \mu \\ \{K_{kM} \ldots K_{(k+1)M-1}\} &= 0, \quad \text{with probability } \gamma \end{aligned} \quad \text{for } \forall k. \quad (10)$$

The definition means that each block of length M is lost with probability γ, and the data loss in different blocks are independent of each other. The power spectral density of the data loss pattern is [3]:

$$S_K(f_k) = \frac{\mu(1-\mu)}{MN} \left| \frac{\sin(f_k \pi M)}{\sin(f_k \pi)} \right|^2 + \mu^2 \delta(f_k), \quad (11)$$

D..3 Markov model-based Data Loss

The Markov model-based data loss is described by the Markov chain shown in Fig. 1. The states of the Markov chain represent the value of the indicator function K_n. If a sample is available at time instant n, the next sample will be available with probability p, and will be lost with probability $1 - p$. If a sample is missing at time instant n, the next sample will be available with probability $1 - q$, and will be lost with probability q. The data availability rate μ is the following [8]:

$$\mu = \frac{q-1}{p+q-2}. \quad (12)$$

The spectral property of a data loss sequence generated by the Markov chain shown in Fig. 1 can be determined according to [8]. The PSD of K_n is a first-order, low-pass type spectrum defined as [3]:

$$S_K(f_k) = \frac{1-a^2}{N(1-a^{2N})} \cdot \frac{1}{|1-az^{-1}|^2} + \mu^2 \delta(f_k), \quad (13)$$

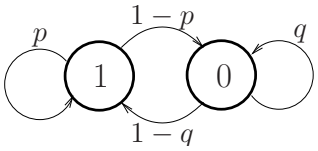


Fig. 1. A two-state Markov model of data loss. State "1": actual sample is available ($K_n = 1$). State "0": actual sample is lost ($K_n = 0$).

where $z^{-1} = e^{-j2\pi f_k}$ and

$$a = p + q - 1. \tag{14}$$

Table 1 summarizes the PSDs of the data loss indicator functions for different data loss models. The small figures in the table illustrate the typical shapes of PSD functions.

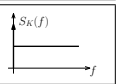
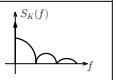
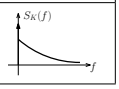
model	$S_K(f_k)$	shape of $S_K(f_k)$
random independent data loss	$\mu^2\delta(f_k) + \dfrac{\mu(1-\mu)}{N}$	
block-based data loss	$\mu^2\delta(f_k) + \dfrac{\mu(1-\mu)}{MN}\left\lvert\dfrac{\sin(f_k\pi M)}{\sin(f_k\pi)}\right\rvert^2$	
Markov model-based data loss	$\mu^2\delta(f_k) + \dfrac{1-a^2}{N(1-a^{2N})}\dfrac{\mu(1-\mu)}{\lvert 1-az^{-1}\rvert^2}$	

Table 1. Summary of PSDs belonging to different data loss models

III. IDENTIFICATION OF DATA LOSS MODELS

Data loss model identification consists of the model selection and the determination of the model parameters. Our previous literature survey has shown that the above three models are appropriate in most cases. The procedure utilizes some direct parameters of K_n and is completed by the evaluation $S_K(f_k)$. The identification process is summarized in Fig. 2.

An essential requirement is that the communication protocol provides information about each sample whether its transfer was successful or not. Without such information only qualitative assessment of the data loss can be done.

If K_n is available, it is also known, whether the protocol is block-based. In the latter case, one sample of K_n is enough for each block representing the data loss. It is a kind of decimation. The block length M is obviously available.

The data availability rate μ can easily be estimated as the mean value or DC level of K_n. This DC level should be subtracted from K_n, in order to remove $\delta(f)$ from the PSD, as its presence can impair the transfer function fitting.

The next step is the calculation of $S_K(f_k)$. It can be

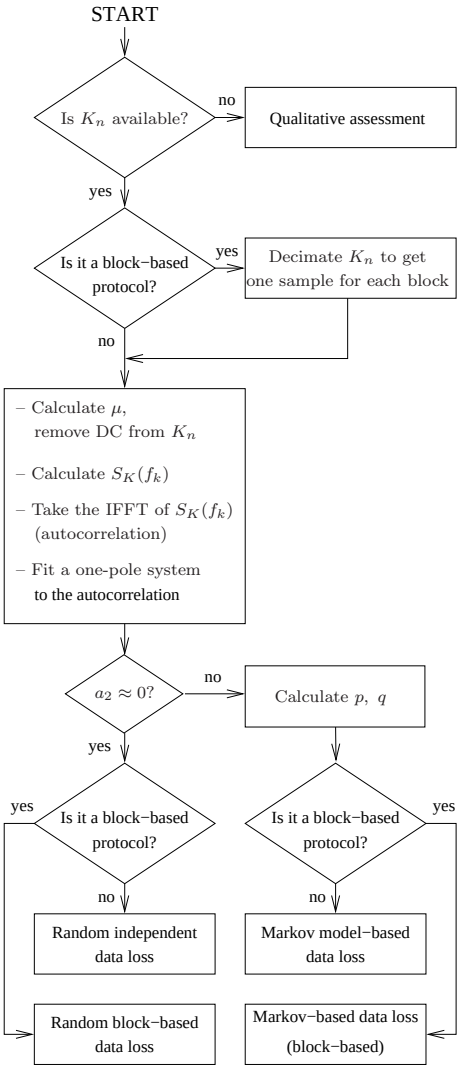


Fig. 2. Data loss model identification

done by the averaging of the FFTs of consecutive (possibly overlapping) blocks of N samples of K_n. As the DC component is removed from the PSD, windowing is not necessary.

The inverse Fourier transform (IFFT) of $S_K(f_k)$ provides the autocorrelation function of K_n. The FFT block size should be greater than the length of the autocorrelation function. It is not a hard requirement, as the usual FFT block size is much greater than required by K_n.

The main part of the identification is the approximation of $S_K(f_k)$. As the previous investigations have shown, the transfer function can be well approached by an all pole or autoregressive (AR) system. Theoretically it has no pole if K_n is random independent, and only one pole if K_n is Markov model-based. $S_K(f_k)$ of a block based data loss has zeros, but after decimation K_n is either random independent or simple Markov model-based. As the system is

quite simple, there are no special requirements. We have used a linear prediction filter (LPC) which determines the coefficients of a forward linear predictor by minimizing the prediction error in the least squares sense [11]. To this end, the `lpc` function of Matlab [10] has been applied.

If already the second LPC coefficient $a_2 \approx 0$, the data loss can be handled as random independent. Its only parameter μ has already been calculated. However, if $a_2 \neq 0$, Markov model-based data loss has been happened. Now the parameters p and q are to be estimated using the relations (12) and (14):

$$\hat{p} = \hat{\mu}(1 - \hat{a}) + \hat{a}, \quad \hat{q} = \hat{\mu}(\hat{a} - 1) + 1, \tag{15}$$

where

$$\hat{a} = -a_2. \tag{16}$$

In Eq. (15) and (16) the hat operator indicates the estimation.

At the end, the information whether the data loss is block based has to be incorporated. If so, the estimators $\hat{\mu}$, $\hat{p}$ and $\hat{q}$ does not change, but the parameter set has to be completed by the block size M.

IV. RESULTS

The procedure presented above has been intensively tested by simulations and measurements. In this section results of both tests are presented. The data processing has followed the procedure given in Fig.2.

A. *Simulation Results*

First a random independent data loss pattern has been generated, then it has been identified by the proposed method. The parameters of the simulation are summarized in Table 2, where L is the total length of the record, N is the FFT size. The spectra have been exponentially averaged, with a smoothing factor α. The constant μ is the parameter of the data loss model. At the end a 10th order LPC model has been fitted, in order to check the dynamic behavior of the data loss. The identified model parameters are the following:

Record length L	FFT length N	smoothing factor α	μ
10^6	1024	0.01	0.9900

Table 2. Main data of the first simulation.

$$\hat{\mu} = 0.9900, \quad \hat{a} = 0.0013, \quad \hat{p} = 0.9900, \quad \hat{q} = 0.0113. \tag{17}$$

As $a_2 \approx 0$, the random independent data loss has been verified. The estimators $\hat{p}$ and $\hat{q}$ are also calculated, and the behavior of the model can also be interpreted by Fig.1.

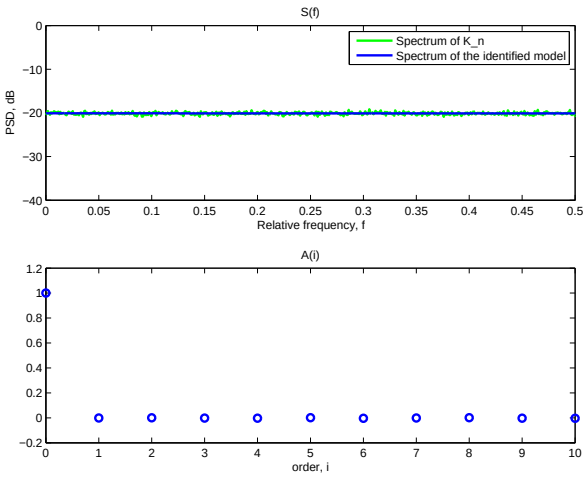


Fig. 3. Identification of a random independent data loss: PSD of the data loss pattern, and the PSD of the model (upper plot). Coefficients of the identified AR system (lower plot).

The system is usually in the '1' state, and if it moves to '0', it has a small probability that stays also in '0' for the next time instant. The PSD of the model has also been calculated. The upper plot of Fig. 3 shows the PSD of the data loss pattern (green line), and the PSD of the model (blue line). The coefficients of the 10th order AR system are depicted in the lower plot. It can be seen that the fitted PSD is in good accordance with the generated one. All the AR coefficients equal approximately zero except the first one.

The second simulation example deals with Markov-based data loss. The parameters of the simulation are summarized in Table 3, where L is the total length of the record, N is the FFT size, and α is the smoothing factor, again. The constants p and q are the parameters of the Markov-model. A 10th order LPC model has been fitted as before. The identified model parameters are the following:

Record length L	FFT length N	smoothing factor α	p	q
10^6	1024	0.01	0.9900	0.9000

Table 3. Main data of the second simulation.

$$\hat{\mu} = 0.9118, \quad \hat{a} = 0.8872, \quad \hat{p} = 0.9900, \quad \hat{q} = 0.8971, \tag{18}$$

The original parameters of the model are p and q, therefore the data availability rate μ is a resulted constant. The second LPC coefficient a_2 is nonzero, but the rest of the coefficients are close to zero. The main result of the identification $\hat{p}$ and $\hat{q}$ are really close to the initial parameters given in Table 3. The PSD of the model has also been calculated. The upper plot of Fig. 4 shows the PSD of the data

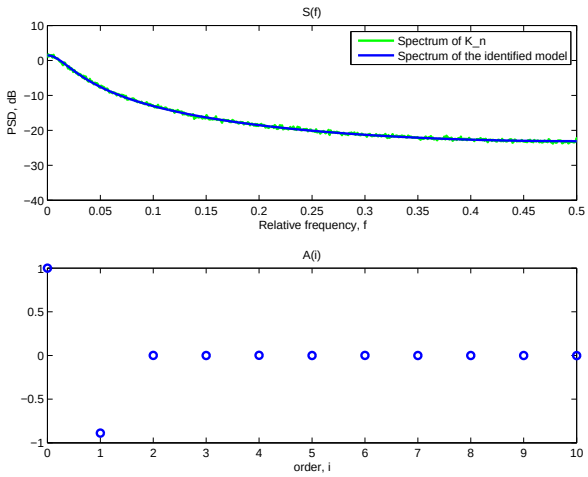


Fig. 4. Identification of a Markov-model based data loss: PSD of the data loss pattern, and the PSD of the model (upper plot). Coefficients of the identified AR system (lower plot).

loss pattern (green line), and the PSD of the model (blue line). The coefficients of the 10th order AR system are depicted in the lower plot. Both the spectra and the LPC coefficients verify that the calculated first order model is appropriate for the Markov-based data loss.

B. Measurement Results

Measurements were carried out by a test system introduced in [9]. In this testbed, wireless sensors perform real-time data collection, and they transmit the data to a PC through a gateway node. In this measurement we used only one sensor. The data sent by the sensor are recorded and processed on the PC. Since data transmission and collection is performed in a hard real-time manner, there is no possibility to apply any acknowledge mechanism for the indication and retransmission of lost packets, hence data loss is inevitable. The data loss is recognized by a time-out mechanism. The sampling frequency is $f_s = 1800$ Hz, and the sensor transmits data in packets of $M = 25$ samples. If data loss occurs, it can be described by the block-based model.

In the first measurement setup, the sensor and the gateway were placed 4 m away from each other in a room, and the sensor was placed near to extensive metal surfaces in order to degrade the radio transmission properties. Depending on the physical circumstances, data loss rates in the range of $[0.1 \ldots 30.0]\%$ could be detected. Now the analysis results for the 3.75% case are introduced. The parameters of the measurement are summarized in Table 4, where L_B is the length of the record in blocks, while t_m is its duration. The parameter N denotes the FFT size, and α is the smoothing factor.

The result of the identification can be seen in Fig. 5. The

Record length in blocks, L_B	Duration of the record, t_m	FFT length N	smoothing factor α
4638	64.4 sec	1024	0.01

Table 4. Main data of the first experiment.

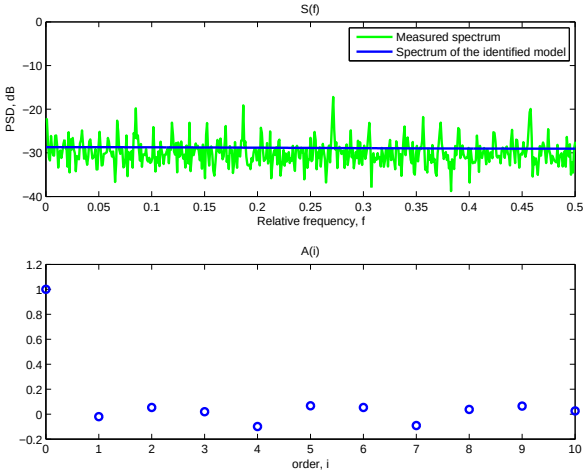


Fig. 5. Identification results of the first experiment: PSD of the measured indicator function, and the PSD of the model (upper plot). Coefficients of the identified AR system (lower plot).

estimated data loss parameters are the following:

$$\hat{\mu} = 0.9625, \quad \hat{a} = 0.0201, \quad \hat{p} = 0.9632, \quad \hat{q} = 0.0569, \tag{19}$$

The PSD of the model has also been calculated. The upper plot of Fig. 5 shows the PSD of the measured data loss pattern (green line), and the PSD of the model (blue line). The coefficients of the 10th order AR system are depicted in the lower plot.

Both the graphical result and the estimated parameters imply that this radio communication suffers from random independent block-based data loss.

The second measurement aimed the investigation of a different data loss mode. A mobile phone has been placed next to the wireless sensor, and the WiFi function of the phone has been activated by playing an on-line media stream. As both devices use the same 2.4 GHz radio band, the communication of the phone causes a disturbance for the wireless sensor. The parameters of the measurement are summarized in Table 5, where L_B is the length of the

Record length in blocks, L_B	Duration of the record, t_m	FFT length N	smoothing factor α
25988	361 sec	1024	0.01

Table 5. Main data of the second experiment.

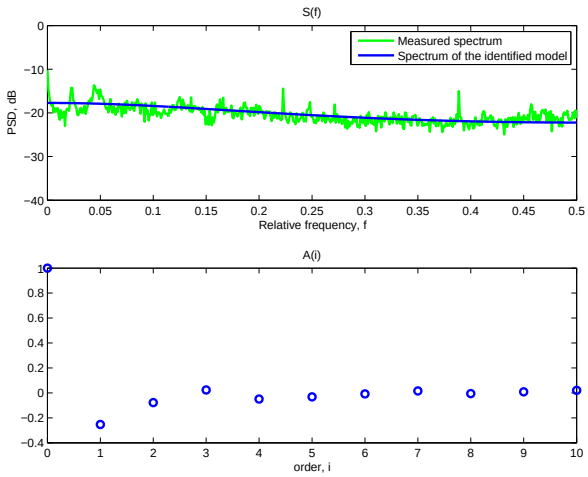


Fig. 6. Identification results of the second experiment: PSD of the measured indicator function, and the PSD of the model (upper plot). Coefficients of the identified AR system (lower plot).

record in blocks, while t_m is its duration. The parameter N denotes the FFT size, and α is the smoothing factor.

The result of the identification can be seen in Fig. 6. The estimated data loss parameters are the following:

$$\hat{\mu} = 0.9533, \quad \hat{a} = 0.2529, \quad \hat{p} = 0.9651, \quad \hat{q} = 0.2878, \tag{20}$$

The PSD of the model has also been calculated. The upper plot of Fig. 6 shows the PSD of the measured data loss pattern (green line), and the PSD of the model (blue line). The coefficients of the 10th order AR system are depicted in the lower plot.

The results clearly show that the data loss introduced by the WiFi function of the mobile phone cannot be random independent. Nevertheless, a Markov-based model can be well fitted to this data loss pattern, as the second AR coefficient is nonzero, while the rest of the coefficients are sufficiently small.

V. CONCLUSION

Recently the analysis of measurement data loss by the spread of sensor networks and Internet-based technology has gained importance. The investigation of the FFT based PSD estimation in the case of data loss has discovered the exact relation between spectral leakage and some data loss models. This paper introduced the inverse procedure: the data loss model can be identified by the Fourier transform of the so-called data availability indicator function. The identification procedure has been elaborated for random independent, random block-based, and Markov model-based data loss. The method has been intensively tested by simulations and measurements. Based on the experiences, the proposed procedure is a promising solution for data loss model identification. Further research is required if the data availability function is not stationary or it is not directly available.

ACKNOWLEDGMENT

This work was partially supported by the ARTEMIS JU and the Hungarian Ministry of National Development (NFM) in frame of the R5-COP (Reconfigurable ROS-based Resilient Reasoning Robotic Cooperating Systems) project.

REFERENCES

[1] L. Kong *et. al*, "Data Loss and Reconstruction in Wireless Sensor Networks," in *Proc. INFOCOM 2013,* Turin, Apr. 14-19, 2013, pp. 1654–1662.

[2] M. Mathiesen, G. Thonet, N. Aakwaag, "Wireless ad-hoc networks for industrial automation: current trends and future prospects," in *Proc. of the IFAC World Congr.,* Prague, Czech Republic, July 4-8, 2005, pp. 89–100.

[3] L. Sujbert and Gy. Orosz, "FFT-based Spectrum Analysis in the Case of Data Loss," *IEEE Trans. Instrum. Meas.*, vol. 65, no. 5, pp. 968-976, May 2016.

[4] J. S. Bendat and A. G. Piersol, *Random Data: Analysis and Measurement Procedures*, New York, London, Sidney, Toronto, John Wiley and Sons, Inc., 1971.

[5] B. Sinopoli, L. Schenato, M. Franceschetti , K. Poolla, M. I. Jordan, and S. S. Sastry, "Kalman Filtering with Intermittent Observations," *IEEE Trans. Autom. Control*, vol. 49, no. 9, pp. 1453–1464, Sept. 2004.

[6] T. Nagayama, B. F. Spencer, G. Agha, and K. Mechitov, "Model-based Data Aggregation for Structural Monitoring Employing Smart Sensors," in *3rd Int. Conf. on Networked Sensing Systems (INSS)*, 2006

[7] O. Hohlfeld, R. Geib, and G. Hasslinger, "Packet Loss in Real-Time Services: Markovian Models Generating QoE Impairments," in *16th Int. Workshop on Quality of Service*, Enschede, June 2008, pp. 239–248.

[8] P. Boufounos, "Generating Binary Processes with All-pole Spectra," in *IEEE Int. Conf. Acoustics, Speech and Signal Processing 2007*, Honolulu, HI, vol. 3, Apr. 15-20, 2007, pp. 981–984.

[9] Gy. Orosz, L. Sujbert, and G. Péceli, "Testbed for wireless adaptive signal processing systems," in *Proc. IEEE Instrumentation and Measurement Technol. Conf.*, Warsaw, Poland, May 1-3, 2007, pp. 123–128.

[10] MATLAB (2010) www.mathworks.com

[11] Jackson, L.B., *Digital Filters and Signal Processing* 2nd ed., Kluwer Academic Publishers, 1989. pp. 255-257.