©2016 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works.

Reference to this paper:

L. Sujbert, Gy. Orosz, "FFT-based Spectrum Analysis in the Case of Data Loss," *IEEE Transaction on Instrumentation and Measurement*, vol. 65, no. 5, pp. 968–976, May 2016.

DOI: 10.1109/TIM.2015.2508278

IEEE website: `ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=7381674`

FFT-Based Spectrum Analysis in the Case of Data Loss

László Sujbert, *Senior Member, IEEE,* and György Orosz

***Abstract*—The significance of measurement data transfer over unreliable channel has emerged in the last decade, due to the spread of sensor networks and the idea of Internet of things. This paper investigates the behavior of the fast Fourier transform (FFT) based power spectral density (PSD) estimation in the case of data loss. There are different methods available to estimate the PSD, but the hegemony of the FFT is beyond dispute, especially in real-time applications. This paper investigates the behavior of the PSD estimator in the case of different data loss models, and then offers some simple solutions on how the data loss can be handled in PSD estimation, when only moderate computing resources are available. The efficiency of the proposed method is demonstrated by the simulation and measurement results.**

Index terms— Data loss, estimation error, FFT, improved estimation, PSD, sensor network.

I. Introduction

Traditional measurement systems provide fast, reliable, and high precision data streaming. However, the technological development in the last decade allowed measurement data transfer in much less reliable systems like sensor networks. In this case, data can be corrupted or the transmission medium can be partially damaged [1], [2]. Recently, the idea of Internet of things has emerged: the connection of physical things to the Internet makes it possible to access remote sensor data and to control the physical world from a distance [3]. The presence of such systems motivates the investigation of data loss phenomena from signal processing point of view. This paper deals with one of such measurement problems, the handling of data loss in the case of spectrum estimation.

Two types of methods can be distinguished according to how missing data are handled in spectrum estimation [4]. In the first approach, missing data are estimated using the existing measurements, and traditional spectrum analysis methods are applied on the reconstructed data set. Missing data reconstruction algorithms are ranging from simple sample-and-hold [5] or slotted resampling [6] techniques to more sophisticated statistical methods [7]. The algorithms can be used either for nonparametric or parametric spectrum estimation like autoregressive (AR) modeling [7].

In the second approach, raw data are processed without any reconstruction. Perhaps one of the most famous work in this field is the Lomb–Scargle method [8], [9]. It can handle even irregular sampling, and produces an estimate of the spectrum by least-squares fitting of sine and cosine components with appropriate orthogonalization technique. The date-compensated discrete Fourier transform (DCDFT) algorithm also uses a kind of sine fitting method on unevenly spaced data [10], and a weight function can be included as well.

Recently, the resonator-based spectral observer (RBO) [11] has been adapted to handle data loss [12]. Unlike the aforementioned methods that operate on an entire data record, it estimates the harmonic components recursively. In [12], the conditions for unbiased harmonic estimation are analyzed. An important result is that the randomness of data loss can guarantee the convergence. The RBO with its basic settings [11] corresponds to the discrete Fourier transform (DFT), which inspired the authors to investigate the behavior of the latter in the case of data loss.

Our literature survey has shown that existing methods deal with spectrum estimation rather than the characterization of distortion caused by missing data. Pinheiro *et al.* [13] and Nagayama *et al.* [14] introduce the bias phenomena caused by data loss, but only qualitative explanations are given. Although [6] deals with bias effects in detail, the effect of data loss patterns on the bias is not considered.

The above-mentioned procedures that can recover missing samples or calculate the spectrum based on the available data require much more computational resources than the wide-spread fast Fourier transform (FFT). The RBO offers some advantages in real-time applications, but it is still too complicated. The hegemony of the FFT is beyond dispute, especially in real-time applications. This paper first investigates the FFT-based power spectral density (PSD) estimation in the case of different data loss models, and then offers some simple solutions how data loss can be handled in spectrum estimation, when only moderate computing resources are available. The results are focused but not restricted to measure harmonic signal components.

This paper is structured as follows. Section II recalls the main steps of power spectrum estimation, and Section III provides a mathematical description of the problem of data loss. Section IV introduces some data loss models accompanied by their spectral features. In Section V a method is proposed that can improve the spectrum estimation in a sense. The results are illustrated by simulation and measurement results in Section VI, while Section VII concludes the paper.

II. Power Spectrum Estimation

The technique recalled in this section is well known, and can be found in many textbooks. The only aim of this overview is

The authors are with the Department of Measurement and Information Systems, Budapest University of Technology and Economics, 1521 Budapest, Hungary (email: {orosz,sujbert}@mit.bme.hu)

to introduce the nomenclature and formulas used throughout the paper. As reference, a basic textbook on random data analysis [15] and an eminent paper on windowing techniques [16] can be cited.

The Fourier transform of a sampled signal $x(t_n)$ can be estimated by a finite set of samples as follows [10]:

$$X(f) = \sum_{n=0}^{N-1} x(t_n) e^{-\jmath 2\pi f_x t_n}. \quad (1)$$

where f_x denotes the real frequency of the signal $x(t)$. From here on, the discrete or relative frequency f is used, i.e., $f = f_x/f_s \in [0 \dots 1]$. The signal $x(t)$ is usually equidistantly sampled, and the spectrum is calculated by the Discrete Fourier Transform (DFT), thus the formula (1) can be rewritten as:

$$X(f_k) = X(k) = \sum_{n=0}^{N-1} x_n e^{-\jmath \frac{2\pi}{N} nk}, \quad n, k = 0 \dots N-1, \quad (2)$$

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

$$S(f_k) = S(k) = \frac{1}{N} |X(f_k)|^2. \quad (3)$$

As the PSD is based on a finite set of samples, it can be calculated even for periodic signals. In the case of non-coherent sampling, the estimation suffers from the phenomena of picket fence and leakage. To suppress these effects, windowing techniques have been developed. Windowing means that the signal x_n is multiplied by the so-called window function w_n prior to the transform:

$$X_w(k) = \sum_{n=0}^{N-1} x_n w_n e^{-\jmath \frac{2\pi}{N} nk}, \quad n, k = 0 \dots N-1. \quad (4)$$

A huge set of window functions has been developed in the last decades. All of them can improve the result of the estimation, and many of them are optimal in a sense.

A significant application of PSD calculation is the analysis of periodic or quasi-periodic signals corrupted by measurement noise. Unfortunately, the measurement noise can hinder the detection of all important harmonic components of the signal. In this case, one finite set of N samples is insufficient, a long series of samples is recorded, many consecutive blocks of N samples are transformed, and the estimator is obtained by averaging the individual PSDs. The blocks can overlap, according to the Welch method. The mean of the individual estimates can be calculated by linear averaging:

$$\bar{S}(k) = \frac{1}{I} \sum_{i=0}^{I-1} S_i(k), \quad (5)$$

where $\bar{S}(k)$ denotes the averaged PSD, and $S_i(k)$ is the PSD of block i. Exponential averaging is also commonly used, when the averaged PSD is calculated in the following way:

$$\bar{S}_{i+1}(k) = \bar{S}_i(k) + \alpha \left(S_i(k) - \bar{S}_i(k) \right), \quad (6)$$

where α is the so-called smoothing constant, $\bar{S}_i(k)$ and $S_i(k)$ denote the averaged and the individual PSD of block i, respectively. For high-precision measurements, the bias caused by the noise can be eliminated by subtracting the PSD of the noise from $\bar{S}(k)$. All the averaging methods require I blocks to provide the spectrum. The length of the record containing I blocks can be treated as the settling time of the averaging.

III. Problem Formulation

A. Formulation of Data Loss

In order to model the data loss, a so-called data availability indicator function, K_n, is introduced [17]:

$$K_n = \begin{cases} 1, & \text{if the sample is processed at } n \\ 0, & \text{if the sample is lost at } n \end{cases}, \quad (7)$$

Samples that are not lost will be termed as processed or available samples. Those DFT blocks that do not contain any lost samples will be termed as complete blocks.

The data loss rate can be defined with K_n as:

$$\gamma = \mathbb{P}\mathrm{rob}\left\{K_n = 0\right\}, \quad (8)$$

where $\mathbb{P}\mathrm{rob}\left\{\cdot\right\}$ stands for the probability operator. The probability that a sample is available is:

$$\mu = \mathbb{P}\mathrm{rob}\left\{K_n = 1\right\} = 1 - \gamma. \quad (9)$$

Data loss rate γ does not determine the distribution of the lost samples in the time domain. A system that is subject to failure can be characterized in the time domain by the reliability function $R(n)$ [24]. $R(n)$ is the probability that the system does not fail in the time interval $(0, n]$. In our framework, failure means that at least one sample is lost in a record, while the reliability equals the probability that no sample is lost. Let the record length be L and the reliability $R(L) = \varepsilon$. Their relationship can be formulated as follows:

$$\mathbb{P}\mathrm{rob}\left\{\prod_{n=1}^{L} K_n = 0\right\} = 1 - \varepsilon. \quad (10)$$

The probability ε is chosen as a small positive number, and the corresponding L defines a record length in which at least one sample is lost with a high probability. For example, $\{\varepsilon = 0.01,\ L = 5000\}$ means that at least one sample is lost within a 5000-sample-long interval with a probability of 99%. The connection between L, ε, and γ depends on the data loss model and will be investigated in the following sections.

B. Spectrum Estimation with Missing Data

Equation (1) is a very simple way of spectrum estimation when the sampling is irregular [10]. Hence, this general form can be easily applied for equidistant sampling and missing data, which is a special case of irregular sampling. Equation (1) implies that if a sample is missing, it is not included in the summation (only existing samples are processed). Using

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

$$X(f) = \sum_{n=0}^{N-1} x_n K_n e^{-\jmath 2\pi f n} = \mathrm{DFT}(x_n K_n). \qquad (11)$$

This formula means that by incorporating K_n into the usual form of DFT, missing samples are practically substituted with zeros. Available samples are weighted with $K_n = 1$, which means no modification. Equation (11) is a very attractive way of spectrum calculation when missing data may exist, since it can be evaluated via FFT.

It is known that (1) often results in biased spectrum estimation for irregular sampling, since the basis functions may not be orthogonal [10]. Since the missing data case is a kind of irregular sampling, and (11) is a special form of (1), a bias can also be expected when (11) is used for spectrum estimation. We will analyze what kind of bias can be expected when (11) is used for spectrum calculation when data are missing, and we propose a simple method that can reduce the bias.

The main idea for the analysis of the bias is that in (11), the signal to be transformed is the product of the lossless signal, x_n, and the indicator function K_n. Hence, the PSD of the signal containing missing samples is obtained as the convolution of the PSD of the lossless signal (S_x) and the PSD of the data loss indicator function (S_K):

$$S(f) = S_x(f) * S_K(f), \qquad (12)$$

where $*$ denotes the convolution operator. The equation shows that a key aspect of the calculation of the PSD of the signal containing missing data is to determine the PSD of the data loss indicator function.

IV. Data Loss Models

In this section, the effect of three basic forms of data loss [5], [18] on the spectrum are 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, and it is often used because of its simplicity [14]. Block-based data loss models are often used, e.g., when several measurement results are transmitted over packet-based communication systems. When a packet is lost, a whole block of data will be missing from the measurement. A real application will be considered in Section VI-B where data loss can be described by the block-based model. Markov data loss models can be used to describe data loss processes when variable lengths of successive measurement samples are randomly missing. Markov model has been proven to be useful, e.g., in the description of data loss pattern in real-time data transmission over the Internet [19].

1) Random Independent Data Loss: Random independent data loss can be defined as follows:

$$\begin{array}{ll} K_n = 1, & \text{with probability } \mu = 1-\gamma \\ K_n = 0, & \text{with probability } \gamma \end{array} \quad \text{for} \quad \forall n. \qquad (13)$$

The definition means that each sample is lost with probability γ, and data losses at different time instants are independent of each other. The time-domain distribution of the data loss is characterized by the $\{L, \varepsilon\}$ couple as defined in (10). The connection to μ can easily be expressed as:

$$\mu = \varepsilon^{\frac{1}{L}}. \qquad (14)$$

The PSD of the data loss pattern is:

$$S_K(f_k) = G + \mu^2 \delta(f_k), \qquad (15)$$

where $\delta(f)$ stands for the Dirac delta function. Since the values of the indicator function, K_n, are independent at different time instants, they are uncorrelated. Hence, the PSD is white, which is represented by the constant term G. The term $\mu^2\delta(f_k)$ represents the power of the DC component (i.e., mean value: μ) of the data loss pattern as given in (9). The calculation of G will be considered in Section IV-A.

2) Random Block-Based Data Loss: To define the random block-based data loss, the time axis should be divided into blocks of length M. The indicator function is given as:

$$\begin{array}{ll} \{K_{kM} \dots K_{(k+1)M-1}\} = 1, & \text{with probability } \mu \\ \{K_{kM} \dots K_{(k+1)M-1}\} = 0, & \text{with probability } \gamma \\ & \text{for } \forall k. \end{array} \qquad (16)$$

The definition means that each block of length M is lost with probability γ, and the data losses in different blocks are independent of each other. The connection between the $\{L, \varepsilon\}$ couple and μ is the following:

$$\mu = \varepsilon^{\frac{M}{L}}. \qquad (17)$$

Note that for a given $\{L, \varepsilon\}$ set, (17) is less than the previous one defined by (14).

The power spectral density of the data loss pattern is [22]:

$$S_K(f_k) = G \left| \frac{\sin(f_k \pi M)}{\sin(f_k \pi)} \right|^2 + \mu^2 \delta(f_k), \qquad (18)$$

The frequency of the occurrence of missing blocks determines the total power of $S_K(f)$, which is included in the term G. The calculation of G will be considered in Section IV-A.

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 [20]:

$$\mu = \frac{q-1}{p+q-2}. \qquad (19)$$

Note that the parameters p, q, and μ cannot be prescribed simultaneously. If the data loss is defined by the $\{L, \varepsilon\}$ couple, the connection to the Markov model parameters can be determined in two steps. First, the probability that no data are lost within an L-sample-long interval is to be expressed. The

probability that the first randomly chosen sample is available equals μ, and the probability that the last $L-1$ samples are not lost equals p^{L-1}. Thus, the required probability is:

$$\mu p^{L-1} = \varepsilon. \tag{20}$$

Suppose that μ is also prescribed, and then p and q are the following:

$$p = \left(\frac{\varepsilon}{\mu}\right)^{\frac{1}{L-1}}, \quad q = \frac{\mu(p-2)+1}{1-\mu}. \tag{21}$$

Actually, the parameters L and ε are completed by the data availability rate μ, and then the probabilities p and q are determined based on this triplet.

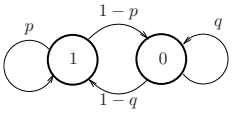


Fig. 1. A two-state Markov model of data loss. State "1": actual sample is available ($K_n = 1$). State "0": actual sample is lost ($K_n = 0$).

The spectral property of a data loss sequence generated by the Markov chain shown in Fig. 1 can be determined according to [20]. Omitting the detailed proof, the PSD of K_n is a first-order low-pass type spectrum defined as:

$$S_K(f_k) = G\frac{1}{|1-az^{-1}|^2} + \mu^2\delta(f_k), \quad a = p+q-1, \tag{22}$$

where $z^{-1} = e^{-\jmath 2\pi f_k}$. Again, the calculation of G will be considered in Section IV-A.

A. Calculation of the Scale Factor

According to the previous section, $S_K(f_k)$ has the general form

$$S_K(f_k) = G \cdot H(f_k) + \mu^2\delta(f_k), \tag{23}$$

where $H(f_k)$ determines the spectral shape of $S_K(f_k)$, and G is an unknown scale factor that can be calculated for discrete PSD as follows [22]:

$$G = \mu(1-\mu) / \sum_{k=0}^{N-1} H(f_k). \tag{24}$$

In the case of block-based and Markov model-based data loss, the scale factor will be calculated using Parseval's theorem:

$$\sum_{k=0}^{N-1} H(f_k) = N \sum_{n=0}^{N-1} h_n^2, \tag{25}$$

where h_n is the impulse response belonging to the PSD $H(f_k)$.

1) Random Independent Data Loss: In this case $H(f_k) = 1$, so the scale factor is:

$$G = \mu(1-\mu)\frac{1}{\sum_{k=0}^{N-1} 1} = \frac{\mu(1-\mu)}{N}, \tag{26}$$

and according to (15) the PSD is:

$$S_K(f_k) = \mu^2\delta(f_k) + \frac{\mu(1-\mu)}{N}. \tag{27}$$

2) Random Block-Based Data Loss: We use the fact that $(\sin(f_k\pi M)/\sin(f_k\pi))$ is the Fourier transform of an impulse train of length M samples, and hence:

$$\sum_{k=0}^{N-1} \left|\frac{\sin(f_k\pi M)}{\sin(f_k\pi)}\right|^2 = N\sum_{n=0}^{M-1} 1^2 = NM. \tag{28}$$

Therefore, according to (18), the PSD is:

$$S_K(f_k) = \mu^2\delta(f_k) + \frac{\mu(1-\mu)}{MN}\left|\frac{\sin(f_k\pi M)}{\sin(f_k\pi)}\right|^2. \tag{29}$$

3) Markov Model-Based Data Loss: We use that the inverse Fourier transform of $(1/(1-az^{-1}))$ is $h_n = a^n$, hence

$$\sum_{k=0}^{N-1} \frac{1}{|1-az^{-1}|^2} = N\sum_{n=0}^{N-1}(a^n)^2 = N\frac{1-a^{2N}}{1-a^2}, \tag{30}$$

Therefore, according to (22), the PSD is:

$$S_K(f_k) = \mu^2\delta(f_k) + \frac{1-a^2}{N(1-a^{2N})}\frac{\mu(1-\mu)}{|1-az^{-1}|^2}. \tag{31}$$

Table I summarizes the PSDs of the data loss indicator functions for different data loss models using the results obtained for the scale factors. The small figures in Table I illustrate the typical shapes of PSD functions.

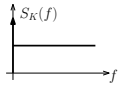
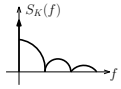
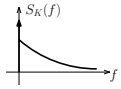
model	$S_K(f_k)$	shape of $S_K(f_k)$
random independent data loss	$\mu^2\delta(f_k) + \frac{\mu(1-\mu)}{N}$	
block-based data loss	$\mu^2\delta(f_k) + \frac{\mu(1-\mu)}{MN}\left\|\frac{\sin(f_k\pi M)}{\sin(f_k\pi)}\right\|^2$	
Markov model-based data loss	$\mu^2\delta(f_k) + \frac{1-a^2}{N(1-a^{2N})}\frac{\mu(1-\mu)}{\|1-az^{-1}\|^2}$	

TABLE I. SUMMARY OF PSDS BELONGING TO DIFFERENT DATA LOSS MODELS.

B. Effects of Data Loss on PSD

1) PSD of a Harmonic Signal: This section discusses what kind of bias is caused by the missing data in the case of a single harmonic signal $x_n = A \cdot \exp(\jmath 2\pi f_0 n)$. The choice is motivated by the fact that all periodic signals can be produced as a superposition of such components. The PSD of $x(t)$ is:

$$S_x(f_k) = A^2 \cdot \delta(f_k - f_0). \tag{32}$$

By performing the convolution as given in (12), one obtains the PSD of a signal corrupted by lost samples:

$$S(f_k) = A^2 \cdot S_K(f_k - f_0), \tag{33}$$

where $g(x) * \delta(x - x_0) = g(x - x_0)$. Equation (33) means that the PSD of the data loss pattern appears around the frequencies where the periodic signal components are located.

Using the general form (23), one obtains:

$$S(f_k) = (\mu A)^2 \cdot \delta(f_k - f_0) + A^2 G \cdot H(f_k - f_0) \qquad (34)$$

The comparison of (32) and (34) shows that two kinds of bias effects can be clearly distinguished:

1) The estimated amplitude of the signal is decreased by factor μ compared with the original amplitude, A.
2) An extra power, a kind of side lobe, appears around the frequency f_0. The power of the side lobe is proportional to A and G, and its spectral shape is determined by $H(f_k)$.

2) PSD of Noise: Let $S_{\mathrm{n}}(f_k)$ denote the original noise spectrum, and $S_{\mathrm{N}}(f_k)$ is its measured PSD in the case of data loss. Using (12), the PSD of the noise corrupted by data loss can be easily computed from the original noise spectrum.

Now, we investigate a practically important case, when the measurement noise is white, i.e., $S_{\mathrm{n}}(f_k) = P$. In this case, using the general form of $S_K(f_k)$ given in (23), the convolution (12) has the form:

$$S_{\mathrm{N}}(f_k) = S_{\mathrm{n}}(f_k) * S_K(f_k) = P \sum_{l=0}^{N-1} (G \cdot H(f_l) + \mu^2 \delta(f_l)) \qquad (35)$$

$$= P \cdot G \sum_{l=0}^{N-1} H(f_l) + P\mu^2 \sum_{l=0}^{N-1} \delta(f_l) = P\mu,$$

where we used the definition of G given in (24). According to the result, the noise level is proportionally changed to the data loss rate, but no extra noise appears due to the data loss.

V. Improved PSD Estimation

A. Proposed Method

An improved PSD estimation technique should avoid the above-mentioned bias and side lobe effects, retaining the resolution of the DFT. In the following, a solution is proposed that requires only moderate extra computation.

Suppose that the PSD is estimated by the averaging of different data blocks. A straightforward idea to avoid the effects caused by data loss is to use only complete blocks, where no samples are missing. If only complete blocks are used for the estimation, all records containing even only one lost sample are thrown away. A question arises how at least a certain part of such records could be used for the estimation.

The idea is the following: find the first lost data position in the block (if there is any lost sample), and fill the rest of the block by zeros. Then this zero-padded block is used for spectrum estimation. The method can be formulated as follows by the redefinition of the availability indicator function:

$$K_n = \begin{cases} 1 & n = 0 \dots n_1 - 1 \\ 0 & n = n_1 \dots N - 1 \end{cases} \qquad (36)$$

where n_1 is the index of the first lost sample in the block. The procedure is demonstrated in Figure 2. Thus, the new spectral block is computed in the following way:

$$X(k) = \frac{N}{n_1} \cdot \mathrm{DFT}(K_n x_n) \qquad (37)$$

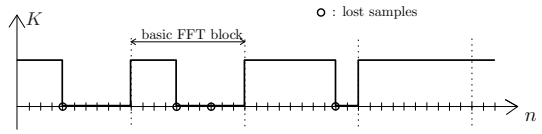


Fig. 2. Modified indicator function of the proposed method.

where N is the length of the DFT. The scaling of the spectrum is necessary to compensate for the lost signal power. Zero padding of the signal samples is a well-known procedure to interpolate the spectrum. Indeed, our proposed method is a kind of interpolation, where the number of the original points is variable, depending on the position of the first lost sample. If $n_1 \ll N$, the side lobe falloff in the spectral block $X(k)$ is very low, compared with the original value. To avoid such a situation, a minimal value $N_{\min}$ of n_1 can be set, and the record is used only if the actual value of n_1 reaches this minimum. As the length of the DFT does not change, the resolution of the spectrum does not change as well.

Figure 2 demonstrates the procedure for nonoverlapping blocks. The efficiency of the method may be further improved if any uninterrupted part of the block (consisting of at least $N_{\min}$ samples) is also used for DFT calculation. However, this would result in overlapped blocks with very short non-overlapping segments, and the noise in the calculated spectral block would not be independent from the previous one, which is useless for the averaging. On the other hand, averaging of overlapped blocks is a common practice in spectrum analysis. In [23], 50% overlap ratio is proposed. The correlation analysis of the overlapping blocks in [16] has shown that an overlap ratio of 75% can further reduce the variance. Based on these results, we propose a maximal overlap ratio of 75%. By this setting, most of available data are utilized for spectrum analysis. According to this value, $N_{\min} = N/4$ is a reasonable setting.

The processing of overlapping blocks is demonstrated in Figure 3. Here $N_{\min} = N/4$ is set, and it equals the length

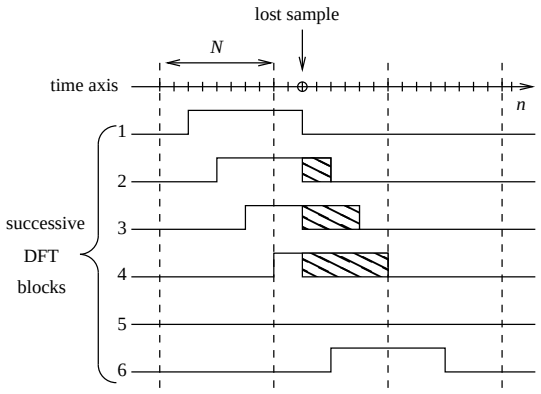


Fig. 3. Processing of overlapped blocks. The striped regions show the zero padding.

of nonoverlapping segments. The first and the last block is

processed in a usual way. Blocks $2 \dots 4$ are processed, but zero padding is necessary. For the fifth one, the nonzero part of the block is too short, and therefore no samples are processed.

B. Settling Times

PSD estimation is ready if enough number of individual PSDs are averaged. As stated before, the simplest solution is to use only complete blocks of N samples, while the proposed method uses all those blocks, for which $n_1 \geq N_{\min}$. It can be supposed that the proposed method needs much less time to collect enough blocks than the straightforward one. As the data loss is random, the faster settling of the proposed method can be shown by the investigation of the probabilities of the occurrence of complete blocks of different lengths.

These probabilities depend in fact on the reliability function $R(n)$. Nevertheless, they can also be expressed by the data loss model parameters, if $R(L) = \varepsilon$ is known. To this end, the relationships in Section IV are used. First, L and ε is set, then the model parameters are calculated, and finally, the probability of complete blocks is expressed.

1) Random Independent Data Loss: The probability of complete blocks of N samples can be expressed by the data availability rate:

$$p_{\text{complete}} = \mu_1^N = \varepsilon^{\frac{N}{L}}, \tag{38}$$

where μ_1 is the data availability rate for this model and expressed by (14).

2) Random Block-Based Data Loss: The probability of complete blocks of N samples (i.e., N/M blocks) can be expressed again by the data availability rate:

$$p_{\text{complete}} = \mu_2^{\frac{N}{M}} = \varepsilon^{\frac{M}{L} \cdot \frac{N}{M}} = \varepsilon^{\frac{N}{L}}, \tag{39}$$

where μ_2 is the data availability rate for this model, and expressed by (17).

3) Markov Model-based Data Loss: It is clear that the reliability function $R(n)$ has an exponential decay for the above two data loss models, and there is a straight relation between the data availability rate (or data loss rate) and the probability of complete blocks. As Markov model-based data loss has two independent parameters, this relation is more complicated. Having the parameters μ and p, the probability can be expressed similarly to (38) and (39):

$$p_{\text{complete}} = \mu p^{N-1}. \tag{40}$$

We have chosen the following setup:

$$\mu = \mu_2, \quad p = \mu_1, \tag{41}$$

where μ_1 and μ_2 are the data availability rates defined for the random independent and the random block-based data loss, respectively. Thus, the data availability rate equals that of the random block-based data loss model. This is reasonable as the Markov-based loss model results in lost blocks as well. On the other hand, assuming that usually $\mu_2 \approx 1$ and $L \gg 1$, the expression of p in (21) and μ_1 in (14) are close to each other:

$$p = \left(\frac{\varepsilon}{\mu_2}\right)^{\frac{1}{L-1}} \approx \varepsilon^{\frac{1}{L}} = \mu_1. \tag{42}$$

	$\varepsilon = 0.01,\ L = 5000$		
	Independent $\mu_1 = 0.999$	Block based $M = 16$ $\mu_2 = 0.985$	Markov model-based $\mu_2 = 0.985$ $p = 0.999,\ q = 0.938$
$N = 256$	$p_{\text{complete}} = 0.790$		
$N = 1024$	$p_{\text{complete}} = 0.389$		
$N = 4096$	$p_{\text{complete}} = 0.023$		

TABLE II. PROBABILITY OF OCCURRENCE OF COMPLETE BLOCKS OF DIFFERENT LENGTHS. DATA LOSS PROBABILITY FOR $L = 5000$ LONG RECORDS IS $1 - \varepsilon = 99\%$.

The probability q can also be expressed by (21). Substituting (41) into (40) we get:

$$p_{\text{complete}} = \mu_2 \mu_1^{N-1} \approx \mu_1^N = \varepsilon^{\frac{N}{L}}, \tag{43}$$

where μ_1 is the data availability rate expressed by (14).

The above analysis has shown that for a given $R(L) = \varepsilon$ the first two data loss models result in exactly equal probabilities of complete blocks of N samples. Having reasonable assumptions, nearly equal probability can be expressed for Markov-based loss model as well. Thus it has been shown that shorter blocks are much more likely to be complete, as it was supposed. Table II illustrates this feature for different data loss parameters. The first row of Table II contains the initial parameters: data loss probability is $1-\varepsilon = 99\%$, and the record length is $L = 5000$ samples. This setting is in accordance with the experimental results introduced in Section VI-B. In the next row, the parameters of the random independent, the random block-based, and the Markov model-based data loss are given. The next three rows show the calculated probabilities for different block lengths. The probability of complete blocks decreases as the block length N increases. In the case of $N = 4096$ only about one block out of 40 could be used for spectrum estimation.

For long records, the expected number of available complete blocks is proportional to the reciprocal of their probability of occurrence. As PSD estimation requires the averaging of a large number of spectral blocks, the proposed method needs less time to settle than the straightforward one.

C. Behavior in the Presence of Measurement Noise

Zero padding in the block can be treated as multiplication by a window function of the length of n_1. The type of the window function is chosen according to the aims of the analysis. Reference [16] discusses the influence of window functions to the measured PSD level in detail. When DFT is applied in the form of (4), the measured power of a harmonic component is multiplied by the square of the sum of the window samples, while in the case of noise, the frequency bins are multiplied by the sum of the squares of the window samples. As a consequence, the level of the resulting PSD depends on the scaling factor N/n_1 defined in (37). It is designed to get the same level for harmonic components, independently from n_1. However, this scaling results in different levels depending on n_1 in the case of noise. [Scaling by $(N/n_1)^{1/2}$ would result in proper measurement of noise, but would deteriorate the measurement of harmonic components.] Since $N_{\min} < n_1 < N$, and n_1

is random for each block, the measured noise level in the averaged PSD is amplified by a random factor in the range of $[1 \dots (N/N_{\min})^{1/2}]$.

D. Computational Complexity

The proposed method calculates the Fourier transform of the blocks of N samples via FFT. It is well known that its computational complexity has the order of $N \cdot \log N$ [15]. The zero padding and the scaling defined in (37) require no extra operations related to N.

The existing methods reviewed in Section I cannot utilize the symmetry of the DFT, as they suppose uneven sampling. The sample-and-hold technique introduced in [5] calculates the correlation function in the discrete time domain, and then the PSD is determined. Because of the time-domain calculation, the computational complexity of the method is in the order of N^2, whatever algorithm is used for PSD computation. The complexity of the AR model fitting [7] depends on the order of the model. To get comparable results, the model order must equal N, as the DFT is able to determine N complex amplitudes, and thus the model requires operations in the order of N^2. The Lomb–Scargle method [8], [9] determines the amplitude and the phase for a single frequency by operations of order N. Assuming again that the PSD is to be determined at N separate frequencies, the method requires order of N^2 computations. Similarly, [10] fits sine wave of a single frequency by the order of N computations, which results in the order of N^2 operations for all frequencies. The RBO [11], [12] works recursively and requires operations proportional to N for one sample in the time domain. It results again a computational complexity of order N^2 for N consecutive samples.

As the above review has shown, the proposed method outperforms the existing solutions, regarding the computational demand. The cited methods are developed to manage different particular problems of uneven sampling and offer solutions that are optimal in a sense. Although the detailed analysis is beyond the scope of this paper, by the above complexity analysis, the FFT-based solution is a real alternative of the referred ones.

VI. Results

The theoretical results derived in the previous sections were verified with intensive simulations. Reference [22] introduced these simulations in detail. First, it has been demonstrated what kind of bias and side lobes occur due to different types of data loss. Then a complex example has shown the viability of the proposed method. In this paper, the measurement results are also presented, while concerning the bias and side lobe effects, we refer to the conference paper [22].

A. Simulation Example

The effects of data loss and the efficiency of the proposed method are demonstrated by a simulation example. The PSD of a signal consisting of two sinusoids is estimated. The signal x_n is the following:

$$x_n = \sin(2\pi f_1 n + 0.2\pi) + 0.001 \cdot \sin(2\pi f_2 n + 0.3\pi), \quad (44)$$

FFT length N	overlap	window type	smoothing factor α	data loss rate γ
1024	no	Hanning	0.01	0.001

TABLE III. Main data of the simulation.

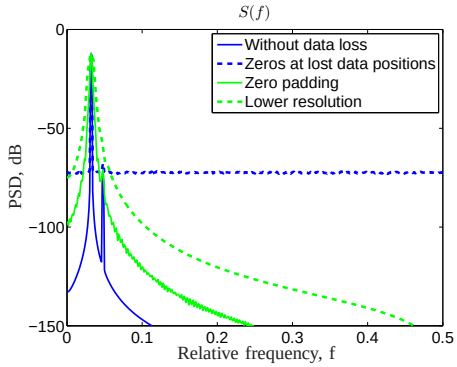


Fig. 4. PSD of the signal given in (44) estimated by different methods.

where $f_1 = 33/1024$ and $f_2 = 49/1024$. The main data of the simulation can be seen in Table III.

The estimated spectra can be seen in Figure 4. The undistorted spectrum is the blue solid curve, while the blue dashed curve belongs to the case when the missing data are replaced by zeros. The spectrum calculated by the proposed method is depicted by the green solid line. The spectrum is also calculated using shorter complete blocks, where the length of the FFT was fixed to $N_{\text{short}} = 256$. The result is the green dashed line. Figure 5 shows the zoomed-in view of the spectral components of the signal. Figure 5 clearly shows that the second sinusoidal of smaller amplitude is difficult to detect if zeros stand for lost samples (blue dashed line) or shorter FFT is calculated from complete blocks (green dashed line). The proposed method allows the detection of the second component as well (green solid line). In addition, there is a frequency mismatch of the low-resolution spectrum (green dashed line), as the peak of the curve appears at a slightly different position.

It can also be seen that the proposed method (green solid

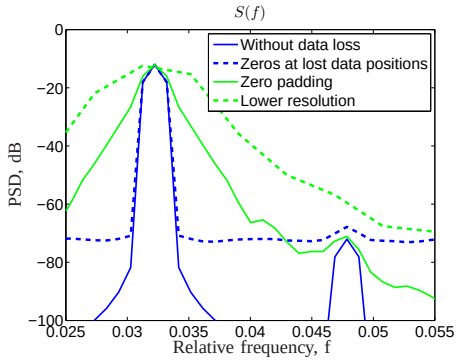


Fig. 5. Zoomed-in view of the harmonic components of the signal given in (44).

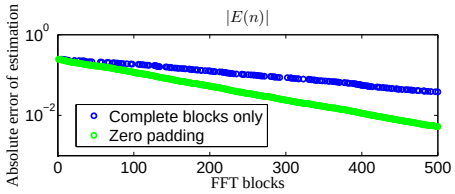


Fig. 6. Settling of two different estimators for the signal given in (44).

line) has greater bandwidth than the simple one replacing the missing data by zeros (blue dashed line). It can hinder the detection of components close to each other. Nevertheless, this feature is similar to windowing, where some advantageous window functions have wider bandwidth than that of the rectangular window.

Another important feature of the proposed estimation algorithm is that its settling is faster. In order to check this, the absolute value of the power estimation is calculated at each step when the estimator is updated. The result for the two important cases can be seen in Figure 6. The settling of the estimator updated based on complete $N = 1024$ long blocks is plotted by blue circles, while the settling of the proposed estimator is represented by green circles. The proposed method outperforms the simple estimator in a convincing manner. The reason is that complete $N = 1024$ long blocks occur with a much less probability than at least $N_{\text{short}} = 256$ long ones as Table II shows.

B. Measurement Results

Measurements were carried out by a test system introduced in [21]. 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 are performed in a hard real-time manner, there is no possibility to apply any acknowledge mechanism for the indication and retransmission of lost packets, and hence data loss is inevitable. The data loss is recognized by a time-out mechanism. The sensor transmits data in packets of 25 samples, and hence the data loss process can be described by the block-based model.

In the 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. In this arrangement, we cannot neglect the presence of data loss.

The measured signal was an amplitude-modulated (AM) signal with a carrier of $f_0 = 100.2$ Hz. The modulation signal was also a sine wave of $f_m = 5.247$ Hz, and the modulation depth was set to 40%. The sampling frequency was set to $f_s = 1800$ Hz, and the transmission was carried out in $M = 25$ long blocks. 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 2.18% case are introduced. Table IV summarizes the settings of the analysis. The measured PSDs can be seen in Fig. 7. The black curve shows the PSD if

FFT length N	overlap	window type	smoothing factor α	data loss rate γ
2048	75%	Hanning	0.02	0.0218

TABLE IV. MAIN DATA OF THE ANALYSIS.

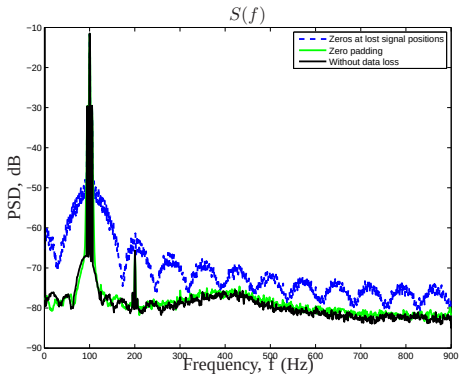


Fig. 7. PSD of the measured signal.

no data are lost. Due to the distortion of the real measurement, second harmonic component and some measurement noise can also be seen in the PSD. The blue dashed curve belongs to the case when the missing data are replaced by zeros. The block-based data loss can easily be recognized by the shape of the spectrum. The PSD calculated by the proposed method is depicted by green solid line. It can be seen that this curve and the original one (plotted by black line) reasonably cover each other, so both the harmonic component and the noise are fairly measured by the proposed method.

Figure 8 shows the zoomed-in view of the spectral components of the signal. The PSD without data loss is not depicted in Figure 8, but the PSD is also calculated using shorter complete blocks, where the length of the FFT was fixed to $N_{\text{short}} = 512$. It is plotted in Figure 8 by the green dashed line. Figure 8 clearly shows the frequency mismatch of the lower resolution PSD and allows to observe the shape of the spectra in detail. Besides the lower side lobes of the proposed procedure, its greater bandwidth is also to be observed. The sidebands are better to be recognized if missing data are

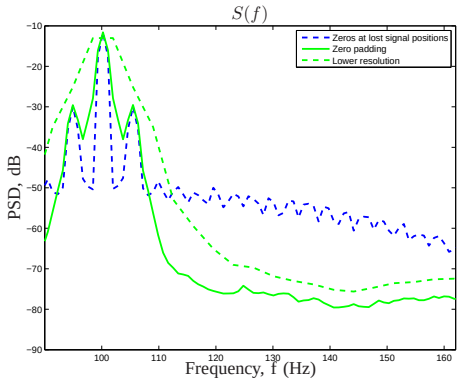


Fig. 8. Zoomed-in view of the harmonic components of the measured signal.

replaced by zeros, but can also be detected by the proposed method. The PSD of shorter blocks does not allow the correct measurement of the AM spectrum.

VII. Conclusion

This paper dealt with the analysis of the FFT-based PSD estimation in the case of data loss. Based on prior work, the behavior of the PSD estimator in the case of different data loss models has been investigated. The bias error of the estimation of harmonic components and the spectral leakage due to different data loss models have been calculated. A simple solution has been proposed when only moderate computing resources are available. The simulations and the measurement results show that our method offers faster settling than the obvious ones, mostly retaining the resolution of PSD calculated only by complete records. Extensive comparison with other methods is left for future research, as well as the refinement of the proposed method.

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] H. Kopetz, *Real-Time Systems*, Springer, 2nd ed. London, U.K.: Springer, 2011, p. 378.

[4] P. Broersen, S. de Waele, and R. Bos, "Application of Autoregressive Spectral Analysis to Missing Data Problems," *IEEE Trans. Instrum. Meas.*, vol. 53, no. 4, pp. 981–986, July 2004.

[5] G. Plantier, S. Moreau, L. Simon, J.-C. Valiere, A. Le Duff, and H. Bailliet, "Nonparametric Spectral Analysis of Wideband Spectrum with Missing Data via Sample-and-hold Interpolation and Deconvolution," *Digital Signal Processing*, vol. 22, no. 6, pp. 994–1004, Dec. 2012.

[6] P. M. T. Broersen, "Five Separate Bias Contributions in Time Series Models for Equidistantly Resampled Irregular Data," *IEEE Trans. Instrum. Meas.*, vol. 58, no. 5, pp. 1370–1379, May 2009.

[7] P. M. T. Broersen, S. de Waele, and R. Bos, "Estimation of Autoregressive Spectra with Randomly Missing Data," in *IMTC2003 - IEEE Instrumentation and Measurement Technol. Conf.*, Vail, CO, USA, May 20-22. 2003, pp. 1154–1159.

[8] N. R. Lomb, "Least Squares Frequency Analysis of Unequally Spaced Data," *Astrophysics and Space Science*, vol. 39, no. 2, pp. 447–462, Feb. 1976.

[9] J. D. Scargle, "Studies in Astronomical Time Series Analysis. III - Fourier Transforms, Autocorrelation Functions, and Cross-Correlation Functions of Unevenly Spaced Data," *Astrophysical Journal*, vol. 343, Part 1., pp. 874–887, Aug. 1989.

[10] S. Ferraz-Mello, "Estimation of Periods from Unequally Spaced Observations," *Astronomical Journal*, vol. 86, pp. 619–624, Apr. 1981.

[11] G. Péceli, "A Common Structure for Recursive Discrete Transforms", *IEEE Trans. Circuits Syst.*, vol. CAS-33, no. 10, pp. 1035-1036, Oct. 1986.

[12] Gy. Orosz, L. Sujbert, and G. Péceli, "Analysis of Resonator-Based Harmonic Estimation in Case of Data Loss," *IEEE Trans. Instrum. Meas.*, vol. 62, no. 2, pp. 510–518, Feb. 2013.

[13] M. Pinheiro, M. Rodriguez-Cassola, P. Prats, A. Reigber, "Analysis of Methods for Reconstructing Periodically Missed SAR Data Acquired Close to Nyquist," in *2012 IEEE Int. Geoscience and Remote Sensing Symp. (IGARSS)*, Germany, Munich, July 22-27, 2012, pp. 307–310.

[14] 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

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

[16] F. Harris, "On the use of Windows for Harmonic Analysis with the Discrete Fourier Transform," in *Proc. IEEE*, vol. 66, no. 1, pp. 51-83, Jan. 1978.

[17] 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.

[18] A. K. Fletcher, S. Rangan, V. K. Goyal "Estimation from Lossy Sensor Data: Jump Linear Modeling and Kalman Filtering," in *Proc. of the 3rd Int. Symp. Information Processing in Sensor Networks*, Berkeley, California, USA, April 26-27, 2004, pp. 251–258.

[19] 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.

[20] 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.

[21] 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.

[22] L. Sujbert and Gy. Orosz, "FFT-based Spectrum Analysis in the Case of Data Loss," in *Proc. IEEE Int. Instrumentation and Measurement Technol. Conf.*, Pisa, Italy, May 11-14, 2015, pp. 800–805.

[23] P.D. Welch, "The Use of Fast Fourier Transform for the Estimation of Power Spectra: A Method Based on Time Averaging Over Short, Modified Periodograms," *IEEE Trans. Audio Electroacoust.*, vol. AU-15, pp. 70–73., Jun. 1967.

[24] A. Høyland and M. Rausand, *System Reliability Theory: Models and Statistical Methods*, 2nd ed., Hoboken, New Jersey, John Wiley and Sons, Inc., 2004.

László Sujbert (S'92-M'95-SM'14) received the M.Sc. and Ph.D. degrees in electrical engineering from the Budapest University of Technology, Budapest, Hungary, in 1992 and 1998, respectively. Since 1992, he has been with the Department of Measurement and Information Systems, Budapest University of Technology and Economics, Budapest. His research interest includes measurement technics, digital signal processing, system identification, active noise control, and embedded systems.

György Orosz was born in Debrecen, Hungary, in 1983. He received the M.Sc. and Ph.D. degrees in electrical engineering from the Budapest University of Technology and Economics, Budapest, Hungary, in 2006 and 2013, respectively. His research interests include digital signal processing, wireless sensor networks, active noise control, and embedded systems.