Bayesian Data Fusion for Pipeline Leak Detection

Marco Guerriero, Fred Wheeler
Analytical Systems Lab
GE Global Research Center, NY, USA

Glen Koste, Sachin Dekate, Niloy Choudhury
Photonics Lab
GE Global Research Center, NY, USA

***Abstract*—In this paper we introduce a probabilistic model for data fusion for leak detection in oil and gas pipelines. We propose a fusion algorithm for both detecting and localizing leaks. Our algorithm optimally combines two heterogeneous systems, fiber optic Distributed Acoustic Sensing (DAS) and Internal Leak Detection (ILD) technology. The output of these two systems, which throughout the paper will be denoted interchangeable as measurements or data points or more frequently as *test statistics* are not necessarily related to physical quantities directly measured by physical sensors. For instance, ILD *virtual sensing* systems, based on computer simulation of pipeline conditions using advanced fluid mechanics and hydraulic modeling, can detect leaks by comparing the measured sensor data (for example flow, pressure or fluid temperature sensors) for a segment of pipeline with the predicted modeled conditions. On the contrary, DAS systems map physical fields acting on the fiber by exploiting coherent optical time domain reflectometry and probing the fiber with proper interrogation systems.**

With ILD we typically have low sensitivity, and poor localization. With DAS we have relatively high sensitivity, but also high false/nuisance alarm rates. By fusing the two we are able to exploit the entirely different physical principle they are based upon, achieving high sensitivity with low false alarms. The fusion process is based on building a Dynamic Bayesian Network (DBN) using the test statistics (which are indicative of a leak) provided by the DAS and the ILD systems. The hidden nodes in the network may indicate the leak/no leak hypothesis the system wants to test and the leak location, respectively. The observable nodes may denote the test statistics from both systems in each bin or zone the pipe is partitioned into. The probabilistic and causal relationships among the nodes are represented and executed as graphs and can thus be easily visualized and extended. The Bayesian perspective of the new analytic allows to easily and naturally incorporate *a priori* information (e.g., wall thickness) on the zone or bin where the leak is most likely going to occur into the leak location node and propagating this new data point through the inference network. Finally we validate our method using computer simulations and real world experimentation conducted by the GE Global Research Center located in Niskayuna, NY. The results demonstrate the benefit of fusing two heterogeneous orthogonal technologies resulting in reduced false alarms, increased response time and improved sensitivity.

***Index Terms*—Bayesian Networks, Generalized Sequential Probability Ratio Test, DAS, fiber optics, SCADA, ILD, pipelines, leak.**

I. Introduction

Today, crisscrossing our world is an underground network of pipelines that deliver oil and natural gas. [1] Spanning more than two million miles, pipes are underneath the streets where we drive, below our rivers and lakes and underneath major businesses. At any given time, various portions of a pipeline may be at risk of malfunctioning, either due to corrosion, mechanical damage, equipment failures, etc. With some pipes 40 years old or more, this infrastructure presents a major challenge for pipeline operators. Their business depends on not just meeting an increasing demand for oil and gas, but also maintaining this network of pipes while operating efficiently, and ensuring the safety of our cities. Just four short years ago, the Bay Area was rocked by a catastrophe [2]. Natural gas inside of pipes deep below the streets ruptured, causing a massive explosion that rocked the community. The losses were devastating, and the repairs, both emotional and physical, are still in process today. The risk of a similar event somewhere in the world is compounded by the sheer complexity, breadth and age of the existing pipeline infrastructure.

It is therefore auspicable to provide systems and methods to provide intelligent pipeline management alarms and/or alerts in an automatic and accurate manner. The motivation to run a Leak Detection System (LDS) [3] on a pipeline is continuous monitoring of system integrity and preparedness for fast initiation of countermeasures in case of a detected and confirmed leakage. Fast and effective mitigation measures can reduce adverse impact on the environment and will substantially reduce the cost for restoration.

Today pipeline operators rely on ILD systems [4]. There are various forms and suppliers, but the general approach is to take in whatever SCADA (Supervisory Control and Data Acquisition) data [5] is available from pressure, temperature, and flow meters, run hydraulic models of the pipeline network, and look for deviations from the expected behavior. ILD systems include balancing methods (where mass or volume flow imbalances indicate a leak) and/or real-time transient-model (RTTM)-based methods, which make it possible to calculate mass flow, pressure, density and temperature at every point along a pipeline in real time - with the help of computational fluid-dynamic model - and provide a fast and sensitive leak-alarm declaration by looking for deviations from the expected behavior determined by the mathematical dynamical model [6], [7]. We can therefore consider the ILD system as a *virtual* sensor, that, given the information available from other measurements and process parameters and by using a set of analytical techniques, is able to measure a quantity of interest, in this case a test statistic indicating the likelihood of a leak. These systems are notoriously inaccurate and generate many false alarms. Moreover they lack the sensitivity to detect leaks below a few $\%$ of flow rate. The localization capability of these systems is limited to the distance between the sensing

points on the pipeline, typically 100 meters if not kilometers. More recently, distributed fiber optic acoustic sensors that measure acoustics, strain or temperature [8], [9], [10] along their entire length of the pipeline are being deployed. DAS are very good at localizing leaks and and have the potential to detect pinhole size leaks and seepage, in addition to detecting third party intrusion such as accidental right-of-way incursion, theft, pipeline contact detection, etc. Their biggest problem is the large number of false alarms.

Apart from systems availability issues, robustness is a function of not relying upon only one physical principle. As an example, if only DAS is deployed on a pipeline, then a rare but specific vibration will affect every single alarm on the pipeline. The same holds true for ILD systems: If only material balance LDSs are used on the pipeline, then any meter failure will affect their performance. Therefore, the deployment of heterogeneous, redundant LDSs is absolutely key in *safety-critical* systems in order to prevent incorrect decisions based on noisy sensor data and ambiguities and inconsistencies present in the environment. Thus a *data fusion* strategy is necessary to exploit data redundancy in order to reduce the effects of the imprecision and noise in the LDS measurements.

The goal of this study is to develop advanced analytics to fuse data from ILD and DAS systems. To this end, we introduce a rigorous mathematical framework where the test statistics produced by the two systems are optimally[1] combined with the prior knowledge of the status of the pipeline from historical records or other information known to the pipeline enterprise (e.g., wall thickness using internal corrosion data). We introduce the following three algorithms:

1) **Algorithm 1**: A batch-based processing fusion algorithm that uses a fixed number of test statistics.
2) **Algorithm 2**: A streaming fusion algorithm that process test statistics streams and it belongs to the class of on-line algorithms with limited memory available to them and also limited processing time per test statistic.
3) **Algorithm 3**: An unsupervised streaming fusion algorithm for which training data under the "leak" hypothesis are not available.

Algorithm 1 implements the optimal Bayesian detector/estimator for the detection of the leak and its localization within the pipe, using a fixed number of test statistics in a batch mode. To reduce the detection response time, Algorithm 2 sequentially process the test statistics produced by the two sensors. Lastly, in the Algorithm 3, we remove the assumption of training the algorithm using "Leak" data and resort to the maximum-likelihood principle to estimate on-line the unknown parameters of the probability distributions of the test statistics, ensuring that the balance between false alarms and sensitivity is active and adjusted automatically to changing conditions.

[1]The algorithm is optimal under the assumptions that the statistical model hold.

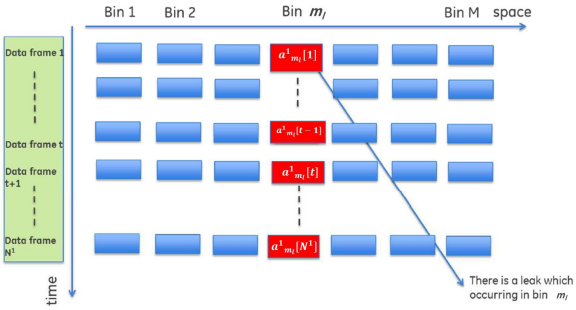

Fig. 1. Pictorial representation of the model of the DAS sensor.

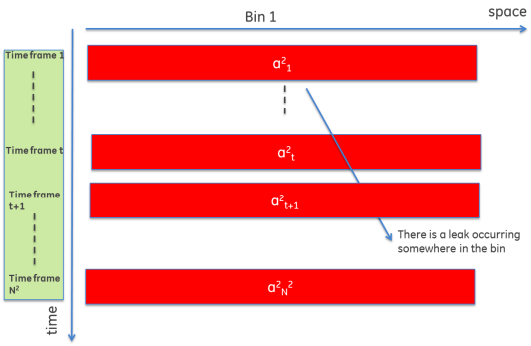

Fig. 2. Pictorial representation of the model of the ILD sensor.

The remainder of the paper is organized as follows. In Section II we formalize our problem. In Section III we introduce Algorithm 1. Sections IV and V describe Algorithm 2 and Algorithm 3, respectively. Algorithm 1 is validated through a combination of simulated and real experimental data in Section VI and, finally, in Section VII we summarize.

II. Problem Formulation and Notations

In general the formulation handles S sensors enumerated by $n \in 1, ..., S$. We describe the system specifically for $S = 2$, with sensor 1 denoting DAS and sensor 2 representing ILD. The Field of Operation (FoO) of each sensor (length of the pipe) is partitioned into segments or bins based on the specific nature of the sensor. For the DAS, the FoO is partitioned into M bins (often called "channels" by DAS suppliers) of equal size (Figure 1). As the ILD does not take any spatial information into account, FoO is simply modeled as one macro bin (2). Therefore ,throughout the paper the bin index m will refer exclusively to the DAS.

Throughout this paper we adopt the following assumptions, definitions and notations

- For sensor 1, the FoO is partitioned into M bins such that each bin m (with $m \in \mathcal{M} = \{1, 2, \ldots, M\}$) has the same size.

- The single leak is located in one and only one bin. We denote this bin with m_l where the subscript l stands for leak.
- The prior probability that the bin m_l contains the leak is denoted with $Pr\{\text{bin } m_l \text{ contains the leak}\} = p_{m_l}$.
- For sensor 1 the measurement $a_m^1[t]$ refers to the acoustic energy in bin m at time t.
- For sensor 2, $a^2[t]$ refers to the mass or volume mismatch at time t in the macro bin [2].
- The measurements $a_m^n[t]$ (also called test statistics) under the $\mathcal{H}_0$ and $\mathcal{H}_1$ hypotheses, are distributed according to the probability density functions (pdfs) $f_0^n(a_m^n[t])$ and $f_1^n(a_m^n[t])$, respectively.
- The measurements $a_m^n[t]$ are independent across the bins and across time.
- To account for time misalignment between the sensors, the number of time frames, which is denoted with N^n, is not the same across sensors.

The measurement data set is given by:

$$\begin{aligned} \mathbf{a} &= \{\mathbf{a}^n[t]\} = \{a_m^n[t]\} \\ n &= 1, 2 \quad \text{sensor number} \\ t &= 1, 2, \ldots, N^n \quad \text{time frame number} \\ m &= 1, 2, \ldots, M \quad \text{bin number} \end{aligned} \tag{1}$$

Measurements with a single subscript refer to all measurements in a single time frame. Measurements with two subscripts identify a specific measurement. It is convenient to define the soft-information likelihood ratio as:

$$\rho_m^n[t] = \frac{f_1^n(a_m^n[t])}{f_0^n(a_m^n[t])} \tag{2}$$

The objective of this study is to develop a new analytic by fusing the test statistics from DAS and ILD, in order to discriminate between the two following hypotheses corresponding to the leak and no-leak events:

$$\begin{aligned} &\mathcal{H}_0 : \text{No Leak} \\ &\mathcal{H}_1 : \text{Leak.} \end{aligned} \tag{3}$$

The performance of detecting $\mathcal{H}_1$ against $\mathcal{H}_0$ is measured by the probability of false alarm and the probability of detection. The probability of false alarm is represented by

$$P_{FA} = Pr(\hat{\mathcal{H}} = \mathcal{H}_1 | \mathcal{H}_0) \tag{4}$$

and the probability of detection is represented by

$$P_D = Pr(\hat{\mathcal{H}} = \mathcal{H}_1 | \mathcal{H}_1) \tag{5}$$

where $\hat{\mathcal{H}}$ represents the detector output.

[2]In order to have mathematical formulations and derivations that are the most general as possible, we introduce the notation $a_m^2[t]$ that is equal to $a^2[t]$ $\forall m$.

III. Batch-based Processing Fusion Algorithm

In this section we develop a window-based algorithm, designed for use in real-time applications, where a subset of the N^n most recent data frames for each sensor n, is used to detect the leak and estimate its location. When a new frame of data is received, the algorithm is repeated after adding the new frame and deleting the oldest frame from the data set.

To discriminate between $\mathcal{H}_0$ and $\mathcal{H}_1$ we propose an optimal (according to a performance metric that captures both the estimation of the bin leaking and the detection) fusion algorithm. The first one is the generalized log-likelihood ratio (GLLR) test which is

$$\begin{aligned} &\text{Accept } \mathcal{H}_1 \text{if } GLLR > \eta \\ &\text{Accept } \mathcal{H}_0 \text{if } GLLR < \eta \end{aligned} \tag{6}$$

where the $GLLR$ is computed as:

$$GLLR = \log \frac{p(\mathbf{a}|\mathcal{H}_1 \text{at bin } \hat{m}_l)}{p(\mathbf{a}|\mathcal{H}_0)} = \sum_{n=1}^{2} \sum_{t=1}^{N^n} \ln \rho_{\hat{m}_l}^n[t] \tag{7}$$

It is not difficult to derive equation (7). In fact from our assumptions and definitions, we can write the likelihood function under $\mathcal{H}_1$ as

$$p(\mathbf{a}|\mathcal{H}_1 \text{at bin } m_l) = \prod_{t=1}^{N^1} f_1^1(a_{m_l}^1[t]) \prod_{m \in \mathcal{M} \setminus \{m_l\}} f_0^1(a_m^1[t])] \prod_{t=1}^{N^2} f_0^2(a^2[t]) \tag{8}$$

In order to compute the estimate of the bin m_l that contains the leak, we adopt the MAP estimator [11] to get:

$$\boxed{\hat{m}_l = \arg \max_{m_l \in \mathcal{M}} p(m_l|\mathbf{a}) = \arg \max_{m_l \in \mathcal{M}} p(\mathbf{a}|\mathcal{H}_1 \text{at bin } m_l) p_{m_l}} \tag{9}$$

It is easy show that the likelihood function given that all measurements are false alarms (there is no leak) is given by:

$$p(\mathbf{a}|\mathcal{H}_0) = \prod_{n=1}^{2} \prod_{t=1}^{N^n} p(\mathbf{a}^n[t]|\mathcal{H}_0) = \prod_{t=1}^{N^1} \prod_{m=1}^{M} f_0^1(a_m^1[t]) \prod_{t=1}^{N^2} f_0^2(a^2[t]) \tag{10}$$

Please refer to Figure 3 for a DBN [12] pictorial representation of this problem. Dividing (8), evaluated at $\hat{m}_l$, by the likelihood in (10), and taking the logarithm of the resulting function, we get the $GLLR$ as per equation (7).

Our first fusion algorithm, which is denoted as Algorithm 1, consists of a cascade of an estimator (which estimates the most likely bin containing the leak) and a detector that distinguishes between the two hypotheses $\mathcal{H}_0$ and $\mathcal{H}_1$ and, furthermore, every time it decides in favor of the $\mathcal{H}_1$, it provides an estimate of the bin containing the leak, that is $\hat{m}_l$. This new type of test was recently proven to be *jointly* optimal in [13], where the word jointly refers to the situation where detection and

estimation strategies have to be solved in a *jointly* optimum manner [3].

Algorithm 1: Batch-based Fusion Algorithm for Leak Detection

1: Set N^1 and N^2.
2: DAS sensor (denotes with the sensor index $n = 1$) acquires samples $a^1_m[t]$ ($\forall m \in \mathcal{M}$ and through $t = 1, ..., N^1$) and computes the MAP estimate $\hat{m}_l$ as per equation (9).
3: Compute $\ln \rho^1_{\hat{m}_l}[t]$.
4: ILD sensor acquires samples $a^2[t]$ (through $t = 1, ..., N^1$) and compute the $\ln \rho^2_m[t] = \ln \frac{f^2_1(a^2_m[t])}{f^2_0(a^2_m[t])}$ which is equal to $\ln \frac{f^2_1(a^2[t])}{f^2_0(a^2[t])}$ $\forall m$.
5: The fusion algorithm computes the $GLLR$ as per equation 7
6: If $GLLR > \eta$, $\mathcal{H}_1$: "leak is present" is claimed; if $GLLR < \eta$, $\mathcal{H}_0$: "NO leak is present" is claimed.

A. *Spillover of leak's acoustic energy to adjacent bins*

The independence assumption of the measurements $a^1_m[t]$ across different bins might be violated in the presence of energy spreading (spillover) into adjacent bins. To model this phenomenon we can use a Markov Random Field [15] where the measurement associated to the leak in bin m_l, that is $a^1_{m_l}[t]$, depends only on the measurements of the adjacent bins. Mathematically eq. (8) becomes:

$$p(\mathbf{a}|\mathcal{H}_1 \text{at bin } m_l) = \prod_{t=1}^{N^1} \left[f^1_1(a^1_{m_l}[t]) \prod_{j \in \mathcal{N}(m_l)} f^1_1(a^1_j[t]|a^1_{m_l}[t]) \prod_{m \in \mathcal{M} \setminus \{\mathcal{N}(m_l) \cup \cup m_l\}} f^1_0(a^1_m[t]) \right] \prod_{t=1}^{N^2} f^2_0(a^2[t]) \tag{12}$$

where $\mathcal{N}(m_l) = \{m_l - 1, m_l + 1\}$ denotes the neighborhood of the bin m_l and $f^1_1(a^1_j[t]|a^1_{m_l}[t])$ the conditional probability of the measurements of the neighbor bins which can be easily derived using the following spillover model:

$$a^1_{m_l-1}[t] = \alpha a^1_{m_l}[t]$$
$$a^1_{m_l+1}[t] = \alpha a^1_{m_l}[t]$$

where $\alpha \leq 1$ is the fraction of energy that spills over the adjacent bins. The corresponding DBN is depicted in Figure 4

[3]Alternatively, one could treat the two sub-problems (detection and estimation) separately and use, in each case, the corresponding optimum technique. In particular, once could use the Neyman-Pearson [14] optimum test for detection and the MAP estimator for the optimum Bayesian estimator as per equation (9) The optimum detection scheme is the well-celebrated likelihood ratio test (LRT), which takes the following form for our specific setup:

$$\mathcal{LR} = \frac{\sum_{m_l=1}^{M} p(\mathbf{a}|\mathcal{H}_1 \text{at bin } m_l) p_{m_l}}{p(\mathbf{a}|\mathcal{H}_0)} \underset{\text{NO Leak}}{\overset{\text{Leak}}{\gtrless}} \eta \tag{11}$$

where $\mathcal{LR}$ denotes the likelihood ratio.

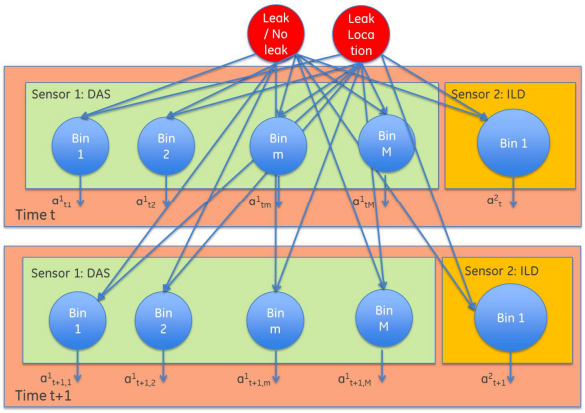

Fig. 3. Dynamic Bayesian Network Structure corresponding to the mathematical representation in (8) and (10).

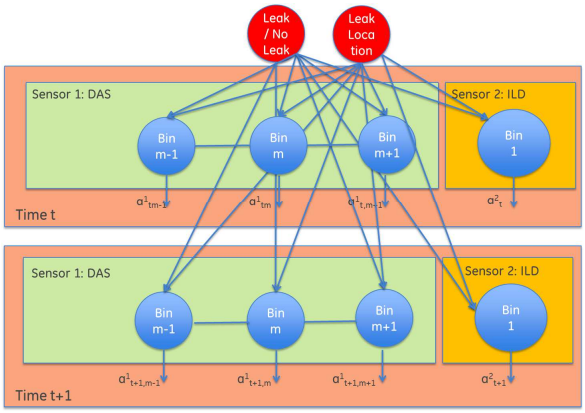

Fig. 4. Dynamic Bayesian Network Structure corresponding to the mathematical representation in (12).

where the edges between adjacent nodes represent the *spillover* effect. It is not difficult to show that the $GLLR$ (which we denote with $GLLR^{spill}$) becomes[4]:

$$GLLR^{spill} = \sum_{n=1}^{2} \sum_{t=1}^{N^n} \left[\ln \rho^n_{\hat{m}_l}[t] + \sum_{j \in \mathcal{N}(m_l)} \ln \frac{f^1_1(a^1_j[t]|a^1_{\hat{m}_l}[t])}{f^1_0(a^1_{\hat{m}_l}[t])} \right] \tag{13}$$

$\hat{m}_l$ is the MAP estimator obtained by using the likelihood function as in (12). The fusion algorithm in this case is the same as the Algorithm 1, provided one replaces the $GLLR$ with $GLLR^{spill}$ and compute $\hat{m}_l$ as per the above discussion.

Further thought reveals that if there is spillover, and if information from contiguous bins is correctly fused, then this spillover of leak energy is actually a boon rather than a nuisance.

[4]The computation of $GLLR^{spill}$ does not really require the introduction of the DBN since we are not fully exploiting the formalism of time via DBN. However, the introduced DBN lends itself to a better extension of the model where i) sensor measurements might be autocorrelated in time, ii) measurements might be statistically dependent from sensor to sensor, iii) the leak flow rate might vary over time according to its own dynamic.

IV. Streaming Fusion Algorithm

To reduce the number of required data frames, which translates into faster leak detection, instead of using a fixed window size $N^n = N$ for all n [5], we can implement a detector for every data frame acquired in a sequential manner by using a Generalized Sequential Probability Ratio Test (GSPRT) [16]. That is, for $N = 1, 2, ...$, we perform the following test:

$$\begin{aligned} &\text{There is a leak in bin } \hat{m}_l \text{ and terminate if } GLLR_N \geq A \\ &\text{There is NO leak and terminate if } GLLR_N \leq B \\ &\text{Take one more data frame if } B < GLLR_N < A \end{aligned} \tag{14}$$

where

$$GLLR_N = \sum_{t=1}^{N} \sum_{n=1}^{2} \ln \rho_{\hat{m}_l}^n[t] \tag{15}$$

where $A > 0$ and $B < 0$ are predetermined constants according to the detection performance objectives:

$$P_{FA} \leq \alpha \text{ and } P_D \geq 1 - \beta \tag{16}$$

Each sensor, computes the log-likelihood ratio for its acquired sample and the fusion algorithm sequentially accumulates the log-likelihood statistics and perform the above test, as described in Algorithm 2.

Algorithm 2: Streaming Fusion Algorithm for Leak Detection

1 Set $N = 0$ and let $GLLR_0 = 0$.
2 **repeat**
3 $N = N + 1$.
4 DAS sensor (denotes with the sensor index $n = 1$) acquires samples $a_m^1[N]$ ($\forall m \in \mathcal{M}$) and computes the MAP estimate $\hat{m}_l$ as per equation (9).
5 DAS sensor computes the $\ln \rho_{\hat{m}_l}^1[N]$.
6 ILD sensor computes the $\ln \rho_m^2[N]$ which is the same $\forall m$.
7 The fusion algorithm updates the sequential $GLLR_N$ according to

$$GLLR_N = GLLR_{N-1} + \sum_{n=1}^{2} \ln \rho_{\hat{m}_l}^n[N] \tag{17}$$

8 **until** $GLLR_N \geq A$ or $GLLR_N \leq B$
9 If $GLLR_N \geq A$, $\mathcal{H}_1$: "leak is present" is claimed; if $GLLR_N \leq B$, $\mathcal{H}_0$: "NO leak is present" is claimed.

Under the assumption that the hypotheses $\mathcal{H}_0$ and $\mathcal{H}_1$ are distinguishable [6], the detection procedure terminates at $N = N_{stop}$. The most beautiful advantage of this sequential detection procedure is the setting of the two thresholds A and B. It can be shown that when P_{FA} and $1 - P_D$ are sufficiently small (that is a desirable objective of any detector) A and B can be written as [16] [7]:

$$A \approx \ln\left(\frac{P_D}{P_{FA}}\right), \; B \approx \ln\left(\frac{1 - P_D}{1 - P_{FA}}\right) \tag{18}$$

V. Unsupervised Streaming Fusion Algorithm

In the previous section we introduced a sequential detection procedure when the pdfs $f_0^n(a_m^n[t])$ and $f_1^n(a_m^n[t])$ under the hypotheses $\mathcal{H}_0$ and $\mathcal{H}_1$ are known. In practice, however, if on the one hand we could expect to leverage a massive amount of data in the no leak (hypothesis $\mathcal{H}_0$) condition, to train the algorithm (estimating $f_0^n(a_m^n[t])$), on the other hand, leak scenarios might be very scarce. This implies that the pdf $f_1^n(a_m^n[t])$ is not exactly known. However, a reasonable assumption is that the family of pdfs under both hypotheses is the same with the complication that under $\mathcal{H}_1$, there normally exist unknown parameters such as the leak signature strength, noise variance to be estimated.

For example, for the ILD, one usually assumes that the pdfs of the test statistic (e.g. flow balance signal) follow a Gaussian distribution with zero mean and standard deviation σ and unknown mean q_l (with q_l indicating the leak flow signal) and the same standard deviation σ which typically captures the precision of the flow meters installed on the pipeline. For the DAS instead, once can legitimately adopt the assumption that the acoustic energies in the frequency range where leaks show their signatures, typically used as test statistics, obey the same probability distribution, but with larger mean under the $\mathcal{H}_1$ hypothesis (with the leak causing an increase in the acoustic energy perceived by the fiber).

Mathematically this can be expressed as the parameters $\theta_1^n \in \Theta_1^n$ under the hypothesis $\mathcal{H}_1$ being unknown, while the parameters $\theta_0^n \in \Theta_0^n$ under the hypothesis $\mathcal{H}_0$ assumed to be known (they can be estimated offline using the enormous amount of no leak available data). With the pdfs $f_0^n(a_m^n[t]; \theta_0^n)$ and $f_1^n(a_m^n[t]; \theta_1^n)$ belonging to the same family of distributions, we further assume that the parameter spaces Θ_0^n and Θ_1^n are disjoint, i.e., $\Theta_0^n \cap \Theta_1^n = \emptyset$. [8]

Since θ_1^n are unknown, we exploit the GLLR test by replacing θ_1^n with their maximum likelihood estimates. The unsupervised streaming fusion algorithm is summarized in Algorithm 3. For the newly acquired measurements $a_m^1[N]$

[5] With a data pre-processing block the two sensors can be synchronized in order to have the same number of measurements at the same time with the same rate.

[6] Mathematically the distinguish-ability of the hypothesis is guaranteed if the second moment of the log-likelihood ratio under both hypotheses is not zero [16].

[7] In a GSPRT the estimate $\hat{m}_l$ introduces an estimation error into the test. In order to compensate for this effect, the thresholds A and B should be function of N, that are A_N and B_N. However, in many practical situations, it is reasonable to assume that $A_N = A + \epsilon$ and $B_N = B + \epsilon$ where ϵ captures the estimation error introduced by $\hat{m}_l$. Therefore, for small estimation error we set constants thresholds as $A_N = A$ and $B_N = B$.

[8] Throughout the paper it is assumed that the parameter spaces Θ_0^n and Θ_1^n are known to the fusion algorithm.

Algorithm 3: Unsupervised Streaming Fusion Algorithm for Leak Detection

Set $N = 0$ and let $GLLR_0 = 0$.

repeat

$N = N + 1$.

DAS sensor (denotes with the sensor index $n = 1$) acquires samples $a_m^1[N]$ ($\forall m \in \mathcal{M}$) and computes the maximum likelihood estimates $\hat{m_l}^{(N)}$ and $\hat{\theta}_1^{1(N)}$ and as per equation (21).

ILD sensor acquires samples $a_m^2[N]$ ($\forall m \in \mathcal{M}$) and computes the maximum likelihood estimates $\hat{\theta}_1^{2(N)}$ as per equation (21). For ILD the $\hat{m_l}^{(N)} = 1$ since there is only one macro bin. Each sensor computes its own $GLLR_{N,n}$ by

$$GLLR_{N,n} = \sum_{t=1}^{N} \ln \left(\frac{f_1^n(a^n_{\hat{m_l}^{(N)}}[t]; \hat{\theta}_1^{n(N)})}{f_0^n(a^n_{\hat{m_l}^{(N)}}[t])} \right) \quad (19)$$

The fusion algorithm computes the $GLLR_N$ according to

$$GLLR_N = \sum_{n=1}^{2} GLLR_{N,n} \quad (20)$$

until $GLLR_N \geq A$ or $GLLR_N \leq B$

If $GLLR_N \geq A$, $\mathcal{H}_1$: "leak is present" is claimed; if $GLLR_N \leq B$, $\mathcal{H}_0$: "NO leak is present" is claimed.

at time N, the parameter estimates (which are $\hat{m}_l$ and $\hat{\theta}_1^n$) are updated according to [9]:

$$\left(\hat{m_l}^{(N)}, \hat{\theta}_1^{n(N)}\right) = \underset{\left(m_l \in \mathcal{M}, \theta_1^n \in \Theta_1^n\right)}{\arg\max} \prod_{t=1}^{N} f_1^n(a^n_{m_l}[t]; \theta_1^n) \times \prod_{m \in \mathcal{M} \setminus \{m_l\}} f_0^n(a_m^n[t]) \quad (21)$$

and the fusion algorithm recomputes the sequential $GLLR_N$ as:

$$GLLR_N = \sum_{t=1}^{N} \sum_{n=1}^{2} \ln \left(\frac{f_1^n(a^n_{\hat{m_l}^{(N)}}[t]; \hat{\theta}_1^{n(N)})}{f_0^n(a^n_{\hat{m_l}^{(N)}}[t])} \right) \quad (22)$$

A. Complexity of the proposed algorithms

If we assume that the batch-based processing fusion algorithm with fixed sample size needs $N^n = N_{fix}$ data points to achieve the detection objectives, its computational complexity is $\mathcal{O}(2N_{fix})$. To achieve the same detection objectives, Algorithm 2 (which is a streaming-based fusion algorithm, has complexity $\mathcal{O}(2N_{fix}^2)$, because $\hat{m_l}^{(N)}$, $\hat{\theta}_1^{1(N)}$, and the $GLLR_N$ need to be computed for every N and the average number of measurements is normally proportional to N_{fix}.

[9]In the proposed sequential algorithm the process of estimating the unknown parameters at each N can be computationally expensive if the stopping rule requires many measurements before reaching the decision. If the estimation process of the unknown parameters can be implemented by recursive algorithms, i.e. simply updating previous estimate as a new measurement is available, it will definitely reduce the computational cost.

B. Example: Gaussian Random Test Statistics with Unknown Means

The test statistics $a_m^n[t]$ for the two sensors, under $\mathcal{H}_0$ and $\mathcal{H}_1$ are assumed i.i.d. Gaussian with mean μ_0^n and μ_1^n and same standard deviation σ_0^n, respectively. Following the assumptions as per before, the means μ_1^n under $\mathcal{H}_1$ are unknown to the fusion algorithm. Assume that

$$0 \leq \mu_1^{n,lower} \leq \mu_1^n \leq \mu_1^{n,upper} \quad (23)$$

where $\mu_1^{n,lower}$ and $\mu_1^{n,upper}$ are the lower and upper bounds μ_1^n.

In the Algorithm 2, equation (19), we have:

$$GLLR_{N,n} = \sum_{t=1}^{N} \left[\frac{\hat{\mu}_1^{n(N)}}{{\sigma_0^n}^2} a^n_{\hat{m_l}^{(N)}}[t] - \frac{{\hat{\mu}_1^{n(N)}}^2}{2{\sigma_0^n}^2} {a^n_{\hat{m_l}^{(N)}}[t]}^2 \right] \quad (24)$$

where

$$\hat{\mu}_1^{n(N)} = \max \left\{ \min \left\{ \frac{\sum_N^{t=1} a^n_{\hat{m_l}^{(N)}}[t]}{N}, \mu_1^{n,upper} \right\}, \mu_1^{n,lower} \right\} \quad (25)$$

VI. Numerical Results

In this section we validate Algorithm 1 assuming the exact statistical models for the test statistics of the DAS and the ILD are known to the fusion algorithm. For the DAS, we use the leak detection test bed that was setup at GE Global Research Center (GRC) in Niskayuna, NY. The test bed consists of three different types of optical cable buried three to six inches under the surface and the trench is $20m$ long. Water leaks were generated using a pressure washer with variable pressures at a distance of $1 - 2m$ from the buried fiber cable and at a depth of $3 - 6$ inches. The water exiting the pressure washer nozzle created an acoustic noise which was detected by the buried fiber using DAS. In the detected signal the acoustic energy within $100Hz$ to $2000Hz$ was higher in case of leak compared to no leak condition. The acoustic energy within $100Hz$ to $2000Hz$ was summed to generate the output test statistic of the DAS. Each leak event was sixty seconds long and each event was broken into sixty events of one second. For each second the DAS output test statistic was computed. Table I lists the leak rate calculated from the water pressure for $1mm$ orifice diameter. It can be seen from the table that leak generated was of the order of 0.85 barrels per hour.

TABLE I
Leak rate for $1mm$ orifice

Pressure (psig)	Leak rate (gpm)	Leak rate (gph)	Leak rate (oil barrels per min)	Leak rate (oil barrel per hour)
400	0.59	35.53	0.014	0.85

Using the real DAS test statistics we are able to learn the pdfs under both hypotheses $\mathcal{H}_0$ and $\mathcal{H}_1$. In Figure 5 the histograms with a Gaussian distribution fit of real DAS test statistics are plotted. For the ILD system, we simulate the test statistics. More specifically, we generated the ILD test

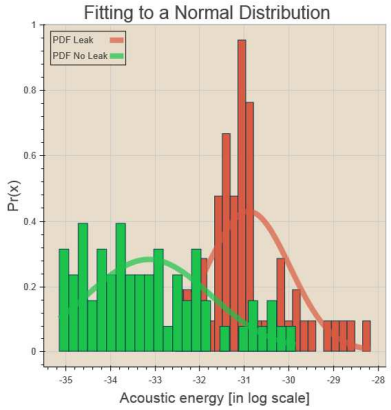

Fig. 5. Histograms with a Gaussian distribution fit of real DAS test statistics for 400 Psi pressure corresponding to a leak rate equal to 0.85 barrels per day.

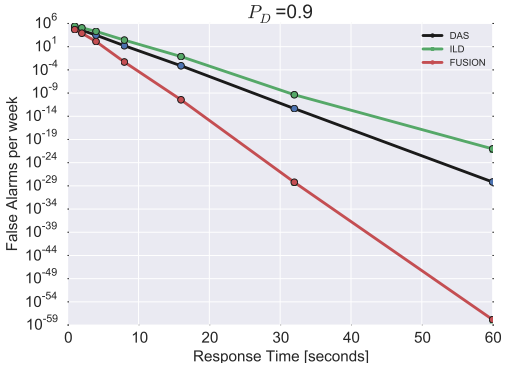

Fig. 6. False alarms per week as function of the response time for a fixed $P_D = 0.9$.

statistics from the Gaussian distribution with means $\mu_0^{ILD} = 0$ ($\mu_1^{ILD} = 1.6$) and standard deviation $\sigma_{0,1}^{ILD} = 1$ under the hypothesis $\mathcal{H}_0$ ($\mathcal{H}_1$). In Figure 6 we show the performance of the Algorithm 1 in terms of false alarms per week versus the response time for a fixed $P_D = 0.9$. Real-world pipeline leak detection systems are usually required to be high-performing, i.e., they must exhibit very low false alarm probabilities. Unfortunately, standard Monte Carlo techniques to estimate the compute the ROC curve become unfeasible for false alarm probability values in the order of 10^{-6} and much lower. To overcome this issue, we used the importance sampling technique as introduced in [17], which can dramatically reduce the number of runs needed to reach a prescribed level of estimation accuracy.

The decrease in false alarms for the Fusion analytic compared to the analytic implemented at the individual systems, is evidently large. For instance, after 8 seconds, the false alarms per week one gets with DAS are 80 vs approximately 500 for ILD. With Algorithm 1, one gets 0.01 false alarms. Moreover, it should be noted that this *fusion gain* is not linear as the observation time increases. Alternatively, one could fix the number of false alarms (that is 0.06 per week) and plot the probability of detection P_D as function of the response time,

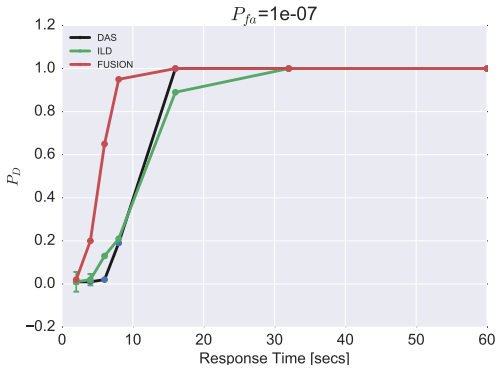

Fig. 7. Probability of Detection as function of the response time for a fixed number of false alarms per week that is $0.0605 = 1e-7 \times 60 \times 60 \times 24 \times 7$.

see Figure 7. Again, there is a large *fusion gain* in the increase in P_D. After 10 seconds, the detection probability for the DAS and the ILD are less than 0.4, while Algorithm 1 already achieves almost complete detectability, that is $P_D \approx 1$.

VII. Conclusions

False alarms are the real pain point for pipeline operators. One usually can reduce them by paying the price of increasing the miss detection rate. With traditional approaches, such as DAS or ILD, the tradeoff between false alarms, detectability and response time is critical. By fusing these two technologies via Algorithm 1, we can alleviate this trade-off by virtually eliminating false alarms and increasing the detectability without sacrificing the response time of the fused detector. These results show a marked increase in sensitivity and in reduced false alarm rated which are the pain point for the pipeline operators. In the future we plan to run and test the streaming algorithms on real field data from ILD and DAS systems in operational pipeline networks.

References

[1] Wikipedia. [Online]. Available: https://en.wikipedia.org/wiki/Pipeline_transport.

[2] ——. [Online]. Available: https://en.wikipedia.org/wiki/2010_San_Bruno_pipeline_explosion.

[3] R. W. Revie, *Oil and Gas Pipelines: Integrity and Safety Handbook*. John Wiley Sons, Inc., 2015.

[4] J. Zhang, A. Hoffman, A. Kane, and J. Lewis, "Development of pipeline leak detection technologies," in *Proceedings of 2014 International Pipeline Conference, IPC2014*, October 2014.

[5] D. J. Gaushell and H. T. Darlington, "Supervisory control and data acquisition," *Proceedings of the IEEE*, vol. 75, no. 12, pp. 1645–1658, Dec 1987.

[6] R. Beushausen, S. Tornow, H. Borchers, K. Murphy, and J. Zhang, "Transient leak detection in crude oil pipelines," in *Proceedings of 2004 International Pipeline Conference, IPC2004*, October 2004.

[7] M. Di Blasi and C. Muravchik, "Leak detection in a pipeline using modified line volume balance and sequential probability tests," *Journal of pressure vessel technology*, vol. 131, no. 2, p. 021701, 2009.

[8] A. Owen, G. Duckworth, and J. Worsley, "Optasense: Fibre optic distributed acoustic sensing for border monitoring," in *Intelligence and Security Informatics Conference (EISIC), 2012 European*, Aug 2012, pp. 362–364.

[9] D. Inaudi and B. Glisic, "Long-range pipeline monitoring by distributed fiber optic sensing," *Journal of pressure vessel technology*, vol. 132, p. 011701, 2010.

[10] M. Nikles, F. Briffod, R. Burke, and G. Lyons, "Greatly extended distance pipeline monitoring using fibre optics," in *Proceedings of OMAE05, 24th International Conference on Offshore Mechanics and Arctic Engineering*, 2005.
[11] H. Van Trees, *Detection, Estimation and Modulation Theory, Part I*. New York, Wiley, 1968.
[12] M. I. Jordan, *Learning in Graphical Models*. New York, Kluwer, 1998.
[13] G. V. Moustakides, G. Jajamovich, A. Tajer, and X. Wang, "Joint detection and estimation: Optimum tests and applications," *IEEE Transactions on Information Theory*, vol. 58, no. 7, pp. 4215–4229, 2002.
[14] H. V. Poor, *An Introduction to Signal Detection and Estimation*. New York: Springer-Verlag, 1998.
[15] R. Kinderman and J. L. Snell, *Markov Random Fields and Their Applications*. American Mathematical Society, 1980.
[16] D. Siegmund, *Sequential Analysis*. New York: Springer-Verlag, 1985.
[17] G. C. Orsak, "A note on estimating false alarm rates via importance sampling," *IEEE Transactions on Communications*, vol. 41, no. 9, pp. 1275–1277, Sep 1993.