2nd International Joint Conference on Water Distribution Systems Analysis & Computing and Control in the Water Industry
Valencia (Spain), 18-22 July 2022
doi: https://doi.org/10.4995/WDSA-CCWI2022.2022.14109

ACOUSTIC DATA ANALYSIS FRAMEWORK FOR NEAR REAL-TIME LEAKAGE DETECTION AND LOCALIZATION FOR SMART WATER GRID

Alvin Wei Ze Chew[1]; Zheng Yi Wu[2]; Rony Kalfarisi[3]; Meng Xue[4]; Jocelyn Pok[5]; Jianping Cai[6]; Kah Cheong Lai[7], Sock Fang Hew[8], Jia Jie Wong[9]

[1, 3,4,5,6] Bentley Systems Singapore Pte Ltd, 1 Harbourfront Pl, Singapore 098633
[2] Bentley Systems, 76 Watertown Rd, Thomaston, CT 06787, USA
[7,8,9]Water Supply (Network) Department, Public Utilities Board (PUB), Singapore

[1] *Alvin.Chew@bentley.com,* [2] *Zheng.Wu@bentley.com,* [3] *Rony.Kalfarisi@bentley.com,* [4] *Meng.Xue@bentley.com,* [5] *Jocelyn.Pok@bentley.com,* [6] *Jianping.Cai@bentley.com,* [7] *Lai_Kah_Cheong@pub.gov.sg,* [8] *Hew_Sock_Fang@pub.gov.sg,* [9] *Wong_Jia_Jie@pub.gov.sg.*

Abstract

Acoustic sensors are widely used for monitoring urbanized water distribution networks (WDNs) to detect and localize pipe leakages. Since their inception, few research studies have focused on developing a generic, effective, and practical methodology to analyse complex acoustics signals for leakage detection and localization in large-scale WDNs. In collaboration with PUB, Singapore's National Water Agency, a generic acoustic data analysis approach has been developed to facilitate PUB's present Smart Water Grid (SWG) management. The proposed approach encompasses multi-stage systematic analyses, namely: (1) data quality assessment; (2) data pre-processing; (3) near real-time leakage event detection and classification; and finally (4) near real-time leakage localization. Our proposed approach is then tested in major WDNs in Singapore having more than 1100km of underground water pipelines and 82 permanently installed hydrophone acoustic sensors between 1 Aug 2019 and 31 Aug 2020, where multiple historical leakage events were reported to within 600m, or less, from neighbouring hydrophones across the large complex networks. By emulating the near real-time detection and localization analyses daily, our proposed methodology could localize reported leakage events to an error range of around 150m on average, while demonstrating significant and stable acoustic leakage power rate over the temporal size of the leakage event cluster(s).

Keywords
Water distribution networks; acoustic signals; leakage detection and localization; acoustic energy analysis; autocorrelation analysis; peaks finding and pairing.

1 INTRODUCTION

Drinkable water is an important resource for humanity's livelihood. With rising uncertainty due to climate change and a growing global population, utility companies are facing increasing challenges to protect and ensure the supply of potable water to the public with minimum disruptions. For example, in the United States, an estimated volume of 6 billion gallons of treated water is reported to be lost each day where approximately 240,000 water mains breaks occur yearly [1]. In the context of Singapore, despite the continual investment in Smart Water Grid (SWG) management [2] by the state government, further reducing non-revenue water (NRW) losses continue to be major challenge due to the complexity of the real-world operations and hidden leakage events which can occur unexpectedly in the underground water distribution networks (WDNs) over time.

Over the last decade, acoustics sensors have been increasingly deployed by utility companies as part of their 24/7 permanently monitoring or temporary leakage program(s) due to low capital cost involved and ease of use. It is believed that acoustic sensors can complement with traditional

flow and pressure sensors to readily detect and localize hidden and insidious leakage events, before becoming disruptive events to the local public. Typically, an acoustic signal, as caused by pipe leakages, is due to the complex interaction between the flowing water and the interiors of the underground pipe wall(s), hence generating random wave signals with both short-term nonstationary and long-term stationary components [3]. Since their inception, many research works have been done to develop different engineering approaches for leveraging on acoustic data signals to perform leakage detection and localization in WDNs which include, but not limited to, traditional experimental analysis to perform signal-based processing [4]–[9], and advanced data-driven and deep learning methods [10]–[12]. Multiple notable works have also been performed using controlled field tests in reasonably large networks (≥120km of pipelines) with high density of acoustic sensors (≥300 sensors) per area [13], [14]. To the very best of our knowledge, while significant research has been done over years, we note that most of the proposed approaches are unlikely to be applicable for the real-world practical context due to the following reasons:

- The conducted works which have achieved high detection and localization accuracies are confined to networks having very high density of sensors per area or pipeline, whether under controlled experimental or field tests. For example, the case studies performed by [13], [14] in the context of Adelaide involved the deployment of high density of acoustic sensors per unit area, in order to detect and localize leakage events to within a limited spatial distance range. However, this requirement may not be generalized to all real-world WDNs as there can be cases having sparse number of acoustic sensors installed permanently, hence there is a strong likelihood that leakage events may occur at reasonably far locations from the nearest acoustic sensor(s), unlike from traditional experiments where sensors are often deployed less than 5-10m away from the simulated leaks in the networks.
- Limited number of acoustic datasets collected in real large-scale systems, pertaining to historical leakage events, available for training detection and localization models via deep learning. In the practical field context, it is not possible to collate large quantity of reported leakage events in well-managed WDNs, thus leading to imbalanced datasets in terms of the total number of leakages to non-leakage acoustic data records for training leakage detection and localization models. This limitation thus challenges the necessity for deploying advanced data-driven and deep learning methods for any engineering modelling objectives, especially in cases having sparse datasets.

To address the above-outlined shortcomings, this work, in collaboration with PUB, Singapore, develops a practically novel and generic acoustic data analysis methodology for analysing complex acoustic data signals collected under uncontrolled field conditions for detecting and localizing historical leakage events in more than 1100km of underground water pipelines with 82 permanently installed hydrophone sensors. This practical setting results in around 1 hydrophone availability for every 15km of pipeline in the combined network that is in stark contrast from the other notable reported field tests [13], [14]. By emulating the near real-time context to analyze acoustic data signals collected from the deployed hydrophones, the proposed methodology comprises of a series of systematic analyses which include: (1) data quality assessment, (2) features generation, (3) data pre-processing, (4) near real-time leakage detection, followed by (5) near real-time leakage localization.

2 DATA DESCRIPTION

Acoustic signal is usually stored as raw audio file which represents the signal's waveform amplitude profile. A typical waveform amplitude profile is illustrated in Figure 1a, which can be converted into its corresponding spectrogram and power spectral density (PSD) profiles as shown

in Figure 1b and 1c respectively. Spectrogram (Figure 1b) comprises of two dimensions where the x-axis represents time (seconds), while the y-axis represents the frequency (Hz) of the acoustic signal. An additional 3rd dimension, as represented by normalized color intensity values in decibels (Db), quantifies the signal strength at a specific frequency value. PSD (Figure 1c) analyses the power density distribution of the same signal over its frequency range, where its x-axis represents the signal's frequency (Hz), and the y-axis represents the corresponding power density (db/Hz) at a specific frequency value.

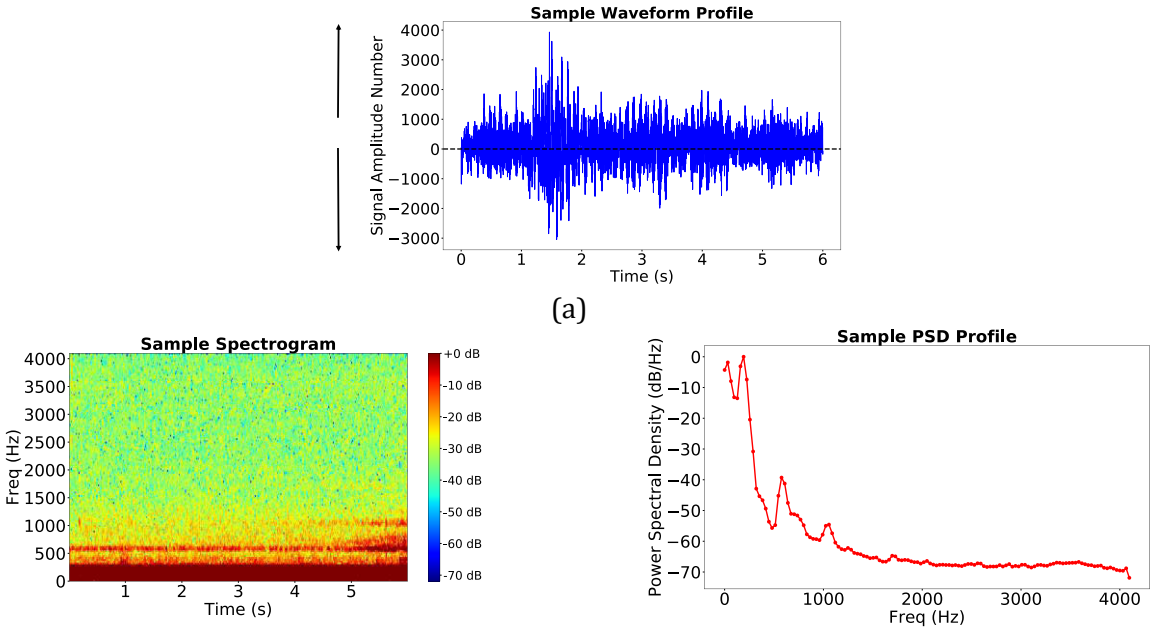


Figure 1. Basics of acoustic signals data: (a) time-series waveform (amplitude) profile; (b) spectrogram; and (c) power spectral density (PSD).

3 METHODOLOGY

Figure 2 summarizes the key systematic procedures involved in our proposed generic leakage detection and localization approach, for every available acoustic sensor station(s) in the WDN system, that comprises of the following components:

i. **Data Quality Assessment:** Remove acoustic audio data files of "bad" quality characteristics.

ii. **Features Generations:** Generate acoustic power features for leakage detection and localization analysis.

iii. **Data Pre-processing:** Remove large transient power values for each acoustic sensor station.

iv. **Near real-time leakage detection and clustering:** Perform outlier detection using pre-processed acoustic power data, followed by clustering the detected outliers into leakage event clusters.

v. **Near real-time leakage localization:** By linking to the detected event clusters, perform leakage localization using autocorrelation analysis for power-peaks finding and pairing.

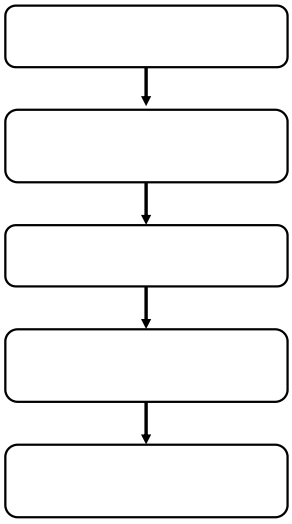


Figure 2. Overview of proposed leakage detection and localization using acoustic signals

3.1 Data Quality Assessment

During the operations of WDNs in the practical field context, a mixture of unknown environmental noises is expected to be embedded in the acoustic signals. It is also common for long-term and permanently installed sensors to not function correctly in the field at all times, hence resulting in numerical errors to be introduced into the recorded acoustic readings over time. Therefore, it is necessary and imperative to assess the initial data quality of each acoustic audio file before further signal analysis. The overall acoustic data quality assessment is performed using several metrics, as summarized in Table 1, while following a proposed screening protocol in Figure 3. An acoustic data file is only classified as "good" quality if it passes all 3 criteria.

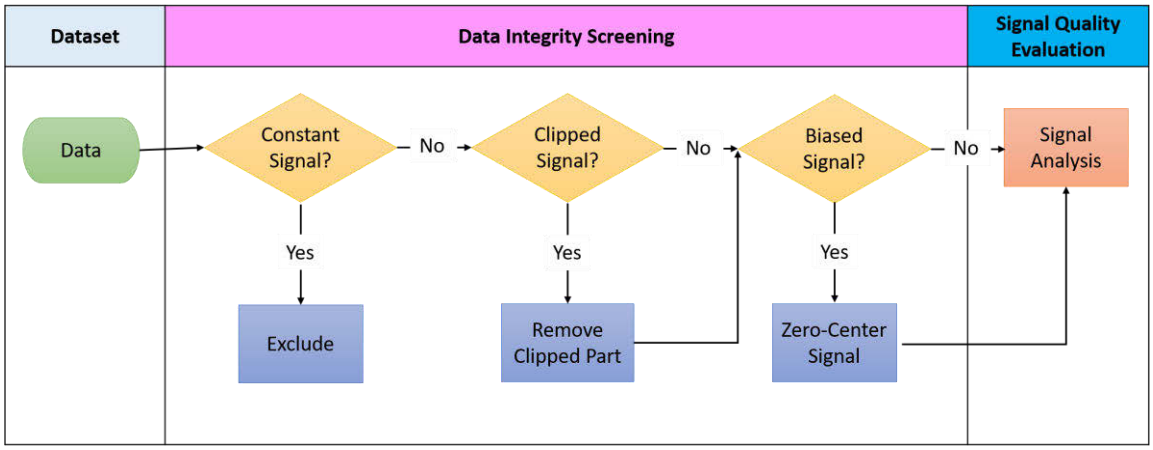


Figure 3. Protocol to systematically preprocess each acoustic data file for constant signals, clipped signals, and offset/biased signal.

Table 1. Descriptions of metrics adopted for performing acoustics data quality assessment.

Data Quality Issue	Problem Descriptions	Rectification measure
Constant Signal	Zero or constant amplitude values for a given datetime	Exclude from analysis
Clipped Signal	Waveform profile is being clipped as only amplitude values within the known upper- and lower bounds for a given bit depth can be recorded	Removal of clipped component
Drifted Signal	Signal amplitude values are not centered along the zero-axis.	Zero-centering of signal values

3.2 Feature Generation

After performing data quality assessment for each acoustic sensor station in the WDN system, we proceed to leverage on the "good" quality audio files to generate their corresponding total power (P_{all}) data representations. Note that P_{all} represents the basic case by using the original waveform amplitude profiles, as previously illustrated in Figure 1a. In general, it is computed by first converting the waveform into its PSD profile (Figure 1c), followed by summing the power density values (normalized or un-normalized form) across their entire or selected frequency range (e.g., 100-750Hz). Note this summation is similar to the traditional power root-mean-square (RMS) computations.

3.3 Data Pre-processing

For each acoustic sensor station, the generated P_{all} values then undergo a series of data pre-processing procedures which encompass the following:

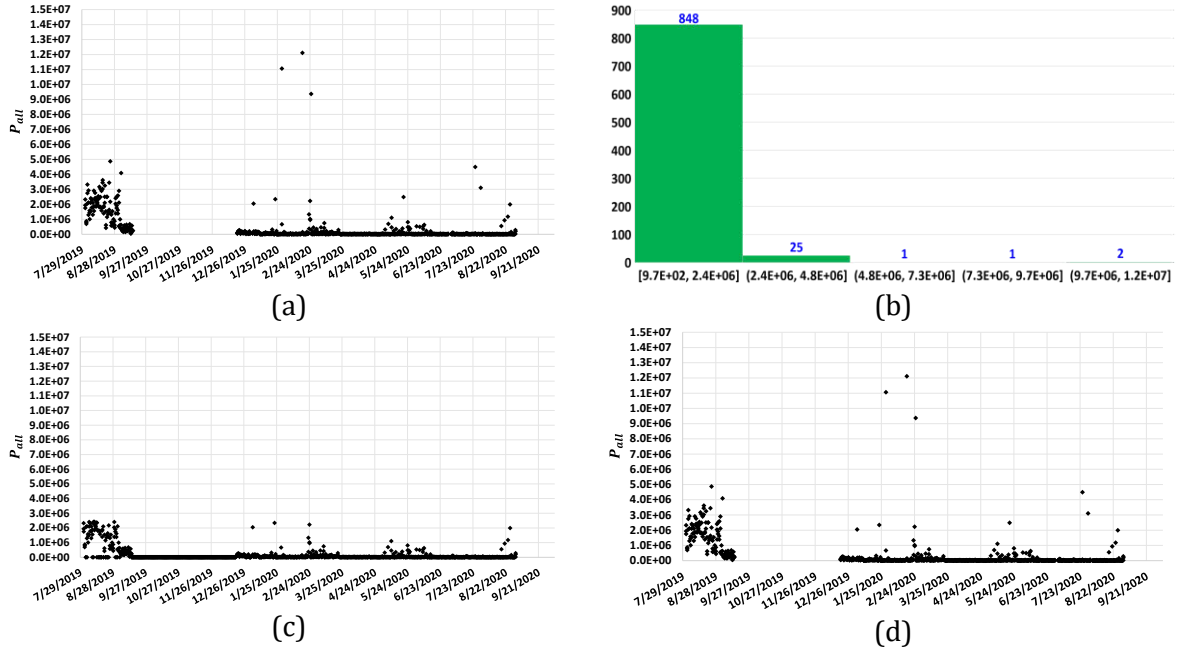
i. **Histogram analysis (1st level of filtering)**: Distribute the data instances with a defined number of bins (N_B) where the data instances from the 1st bin are retained and the remaining data instances from the other bins, i.e., 2nd bin and beyond, are collated together.

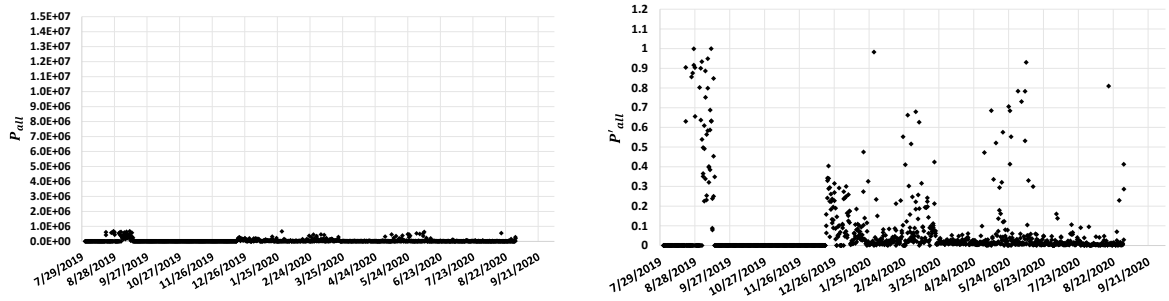
ii. **Data restoration**: For each of the data instances in the combined 2nd bin and beyond, compute the ratio between available adjacent data instances, namely: (1) between $P(t)$ and $P(t-1)$; and (2) between $P(t)$ and $P(t+1)$, where $P(t-1)$ and $P(t+1)$ are the original data instance values. If either of the computed ratio values from the above (1) and (2) computations are within a defined threshold scaling value (S_{thres}), $P(t)$ data instance will be formally restored back into the original 1st bin of data instances, else $P(t)$ will formally remain in the 2nd bin and beyond. At this stage, the original 1st bin of data instances has been updated with any restored data instances. Finally, for the data instances, in the 2nd bin and beyond, without any adjacent neighboring values, they will formally remain in the 2nd bin and beyond.

iii. **Histogram re-analysis (2nd level of filtering)**: Re-distribute the data instances in the final 1st bin of data instances from the preceding step (iii) with the same number of bins (N_B), as previously defined in step (i), to obtain a new 1st bin of data instances.

iv. **Data normalization**: For the new 1st bin of data instances after the data re-distribution from preceding step (iii), normalize each data values accordingly (e.g., max-normalization or min-max-normalization).

To illustrate the above-proposed data pre-processing procedures, Fig. 4a shows a typical time-series profile for P_{all}. Fig. 4b plots the distribution of the power values across 5 bins (N_B). Fig. 4c shows the collated data instances (total of 848) from the 1st bin as derived from Fig. 4b, while Fig. 4d illustrates the updated time-series profile after data restoration. By performing another round of histogram re-analysis using the collated data values from Fig. 4d, Fig. 4e plots the time-series profile for the final 1st bin of data values. Finally, Fig. 4f illustrates the normalized power values from Fig. 4e by using max-normalization method.





(e) (f)

Figure 4. Example for pre-processing procedures for P_{all} values: (a) original time-series profile; (b) histogram distribution for original profile; (c) time-series profile after 1st filtering; (d) time-series profile after data restoration; (e) final time-series profile after 2nd filtering; (f) normalized final time-series profile

3.4 Near real-time leakage detection & clustering

Using their corresponding normalized time-series profiles for P_{all}, leakage detection and classification is then performed for each available acoustic sensor station in the system by adopting temporal-based clustering which depends on several key model parameters, namely: (a) universal reference value (P_{ref}) to perform power-based outlier detection; $0 \leq P_{ref} \leq 1.0$, (b) minimum number of detected outliers ($N_{outlier}$) between 2am-4am required for each station daily; $1 \leq N_{outlier} \leq 3$; (c) minimum number of consecutive days (N_{cons}) for each station to form an anomaly cluster where each of the days fulfilled the prior $N_{outlier}$ value defined; $1 \leq N_{cons}$.

For each selected combination of P_{ref}, $N_{outlier}$, and N_{cons} (from model training) in the near real-time context, the following set of systematic analyses is performed for each acoustic sensor station.

i. **Power-based outlier detection:** Compare each of the normalized data values with the universally defined P_{ref} value. If the normalized data value is greater than P_{ref}, then the corresponding timestamp is marked as a detected outlier.

ii. **Leakage Event Clustering:** The detected outliers are then aggregated along the time horizon and subsequently clustered or classified into leak events, where each of the detected events is required to fulfill the following criteria:

   a. **Basic Criterion 1 (BC-1)**: The total number of outliers detected during the MNF hours (e.g., 2am-4am) daily ≥ $N_{outlier}$.

   b. **Advanced Criterion 1 (AC-1):** The detected outliers are then aggregated over consecutive number of days into a single anomaly cluster event, where its corresponding size ($S_{current}$) ≥ N_{cons}.

iii. **Near real-time analysis:** Referring to Figure 5, on a daily basis, **BC-1** must be first fulfilled, followed by adding the identified number of outliers into $S_{current}$. If **BC-1** is not fulfilled on any given day, **AC-1** is then triggered to check if the respective criterion is fulfilled for $S_{current}$. $S_{current}$ is then reset to 0 on the following day for continuing the near real-time analysis.

Note that the selected P_{ref}, $N_{outlier}$, and N_{cons} parameters for the near real-time analysis are usually determined/optimized from the model training phase. Figure 6 exemplifies our proposed leakage event detection and clustering by using $P_{ref} = 0.4$, $N_{outlier} = 1$, and $N_{cons} = 3$, which results in $S_{current}$ to be 11 days where no outliers are detected after Day 11th as shown.

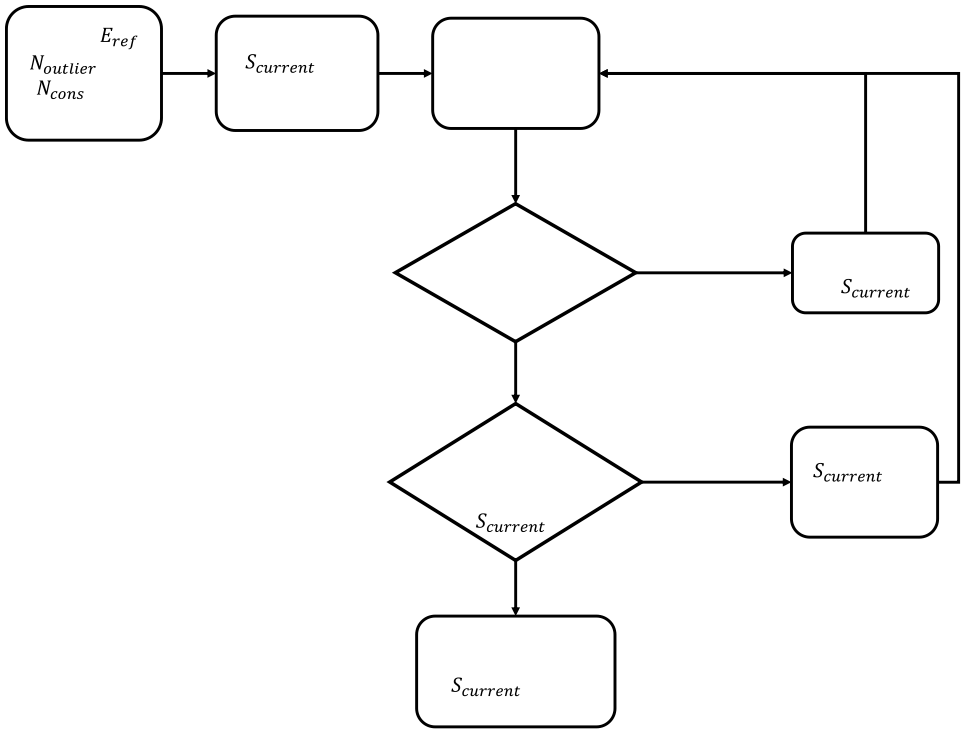


Figure 5. Proposed procedures for performing near real-time leakage detection and clustering.

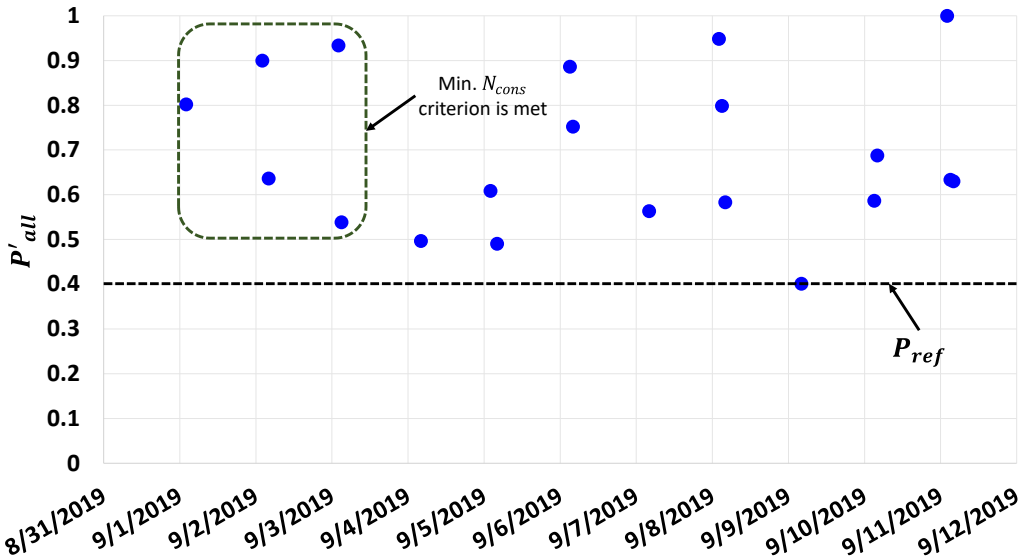


Figure 6. Example for anomaly event detection and clustering using $P_{ref} = 0.4$, $N_{outlier} = 1$, and $N_{cons} = 3$.

3.5 Near real-time leakage localization

Upon detection of leakage event clusters based upon any selected combination of P_{ref}, $N_{outlier}$, and N_{cons}, leakage localization is performed with the corresponding audio files for the associated detected outliers within the event cluster in the same near real-time context by adopting a set of mathematical procedures as follows:

i. **Bandpass filtering:** For each detected outlier that corresponds to a specific audio file, apply bandpass filter of a defined frequency range (e.g., 100-750Hz) to its original

waveform profile to extract a filtered waveform profile, followed by generating its corresponding spectrogram.

ii. **Averaged spectral amplitude:** For the same selected frequency range from (i), derive the averaged spectral amplitude profile from the generated spectrogram using a defined number of moving average points (N_{avg}).

iii. **Autocorrelation analysis:** Apply autocorrelation function to averaged spectral amplitude profile that analyses all possible time-lags for the total time-length of the audio file.

iv. **Power-peaks finding & pairing:** Perform peak finding and pairing on the derived autocorrelation plot from (iv) using suitable confidence intervals (e.g., 95%) for multiple tolerance bounds as follows:

   a. **Horizontal upper- and lower-bounds for autocorrelation values:** $\pm\frac{Z}{\sqrt{N}}$, where N represents the total number of data points in the averaged spectral amplitude time-series profile, and Z represents the t-statistic score for a defined confidence interval. In principle, the autocorrelation values which are outside of the upper- and lower-bounds are retained for the subsequent analysis.

   b. **Vertical bound for time-lag values:** For the same defined confidence interval, estimate the vertical time-lag bound using $\mu + n\frac{\sigma}{\sqrt{M}}$, where μ and σ represent the harmonic mean and standard deviation of the time-lag values corresponding to the M number of retained autocorrelation values from (a), and n refers to the total number of sigmas to be considered. For example, for 95% confidence interval, n equates to 3.

v. **Localization distances estimation:** For each pair of the power peaks identified from (iv), check if (1) their corresponding time-offset (Δt) is within a defined threshold time-offset (T_{thres}), and (2) the corresponding absolute difference between the pair of autocorrelation values is within a defined threshold autocorrelation tolerance (A_{thres}). If both criteria (1) and (2) are fulfilled, proceed to estimate the localization distance (d_{leak}) for each pair by multiplying Δt with the speed of sound in water (v_{sound}). Collate and distribute all estimated d_{leak} values via histogram analysis with a defined distance width (d_{width}), followed by determining the average localization distance (D_{leak}) from the specific bin having the highest frequency count.

Figure 7 exemplifies the key procedures involved to estimate the localization distance(s) for a singular detected outlier and its associated audio file, where Figure 7a represents the original and bandpass filtered waveform profile for the audio file. Figure 7b then illustrates the averaged spectral amplitude profile derived from the bandpass filtered waveform, followed by using the filtered waveform to generate its autocorrelation plot and the associated peaks as shown in Figure 7c. The same figure illustrates the resulting upper- and lower-horizontal bounds using 95% confidence interval to first identify the statistically significant peak autocorrelation values, i.e., those outside of the two bounds, followed by plotting the vertical bound, as shown, to isolate the key peak values to within a certain time-lag (s). Finally, Figure 7d shows the histogram plot for 3 estimated d_{leak} of 40.5m, 445.2m, and 829.6m for 3 unique pairs of peaks identified from Figure 7c, where each distance has a count of 1 when d_{width} is fixed at 50.0m and v_{sound} = 1480m/s. In cases where there are multiple dominant localization distance ranges having the same counts, then we estimate their harmonic mean value among them, which will result in a D_{leak} value of 106.6m for the current selected example.

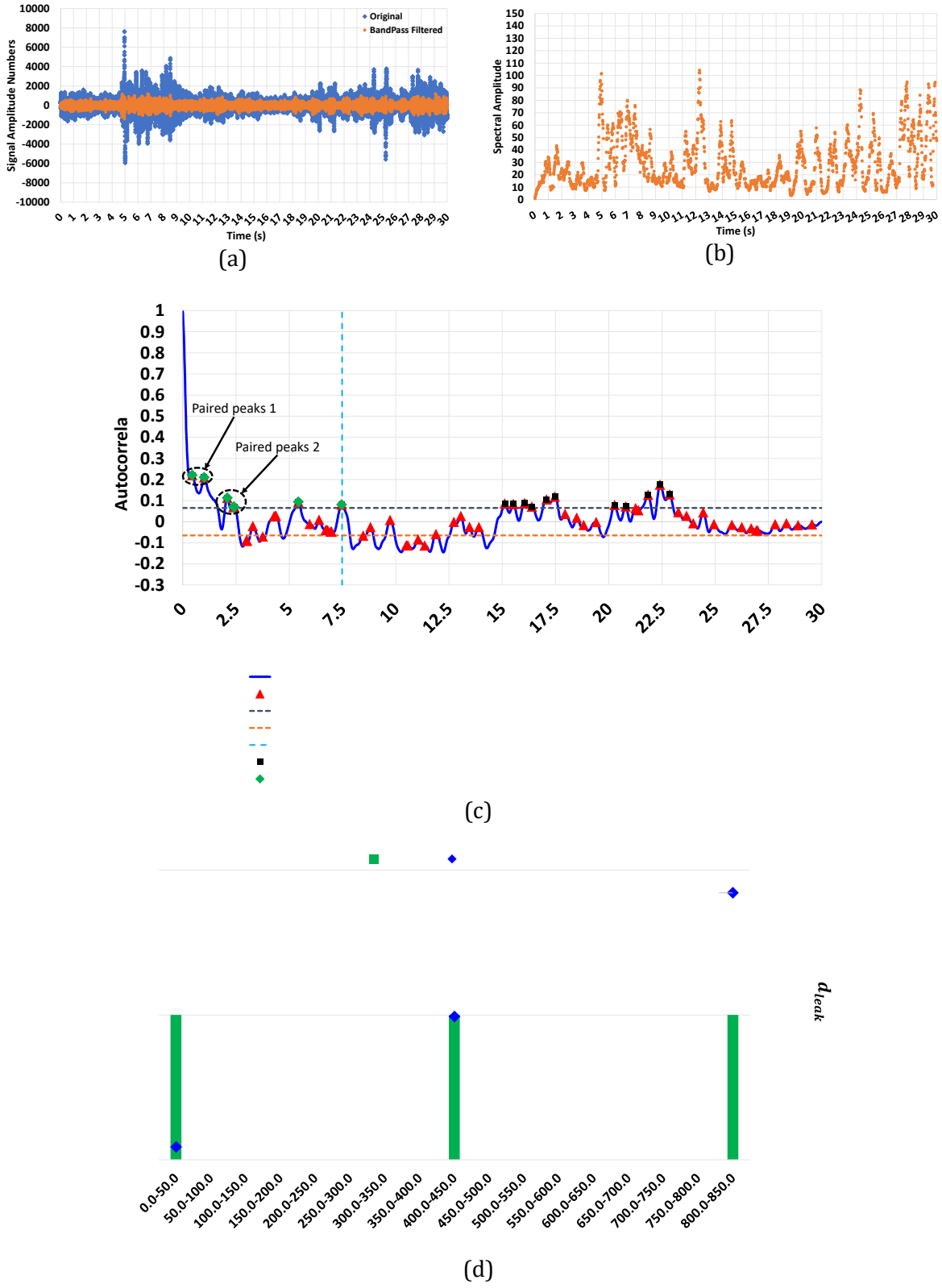


Figure 7. Example for estimating localization distance for each detected outlier: (a) applying bandpass filter to original waveform profile; (b) derivation of averaged spectral amplitude profile from filtered profile; (c) derivation of autocorrelation plot with paired peaks; (d) localization distances estimation and histogram analysis

4 CASE STUDY

4.1 Description of WDN systems

In collaboration with PUB, Singapore's National Water Agency, our proposed leakage detection, and localization methodology is verified with large-scale WDNs, which encompass three water supply zones (Zone-1, Zone-2, Zone-3) in Singapore, as shown in Figure 13. All three zones consist of underground water pipes having a total length of 1100km and 82 permanently installed hydrophone sensors. Table 2 summarizes the details of the audio files collected across all hydrophones in all three zones for the period between 1 Aug 2019 and 31 Aug 2020, where 18 historical leakage events were reported to within 600m, or less approximately, from neighboring hydrophone acoustic sensors, as summarized in Table 3 for the respective zones.

Table 2. Details pertaining to acoustic data files collected in Zone-1, Zone-2, and Zone-3.

Detail	Zone-1	Zone-2	Zone-3
Date range		1 Aug 2019 – 31 Aug 2020	
Total quantity of hydrophones	27	47	8
Total quantity of .WAV files	156734	232026	44374
Total quantity* of .WAV files from MNF hours	21734	25518	6109
Bit depth of .WAV files		16	
Sampling rates (Hz)		2048-8192	
Time length of .WAV files (s)		6.0-30.0	
No. of channels		1 (mono)	

**** After undergoing data quality assessment***

Table 3. Summary of reported leakage events in Zone-1 to Zone-3, and their associated nearest hydrophone stations.

Zone	Reported Leak Dates	1st detected date (nearest leakage cluster to reported leak)	Nearest station	Pipeline distance between leak & nearest station (m)	No. of detected leakage clusters	Avg. Predicted localization dist. (m)
1	9/11/2019	9/1/2019	STN_A	72	2	174
	9/9/2019	9/5/2019	STN_B	419	12	263
	8/16/2019	8/12/2019	STN_C	587	12	517
2	1/9/2020	1/5/2020	STN_A	36	3	157
	5/22/2020	5/20/2020	STN_B	111	7	363
	8/1/2019	8/2/2019	STN_C	202	3	234
	8/5/2020	8/3/2020	STN_D	324	7	146
	5/6/2020	5/7/2020	STN_E	530	7	473
	3/13/2020	3/7/2020	STN_F	600	4	280
	3/12/2020	3/7/2020	STN_F	600		280
3	3/6/2020	3/2/2020	STN_A	353	9	654
	1/23/2020	1/24/2020	STN_B	453	19	415
	3/28/2020	3/24/2020	STN_C	518	13	450
	6/28/2020	6/24/2020	STN_C	627		514
	1/21/2020	1/17/2020	STN_D	613	21	625
	1/20/2020	1/17/2020	STN_D	736		625

4.2 Near real-time leakage detection and localization

4.2.1 Leakage event detection

By emulating the near real-time context using the historical reported leakage events, we first perform the leakage detection and clustering analysis by adopting the model parameters of $P_{ref} = 0.4$, $N_{outlier} = 1$, and $N_{cons} = 3$. Table 4 summarizes the total number of detected event clusters for the respective station in each zone, and the 1st detected date of the event cluster located closest to the reported event temporally. For example, for STN_A in Zone-1, there are 2 detected event clusters for that station during 1 Aug 2019 and 31 Aug 2020, where the detected cluster nearest to the reported event on 9/11/2019 is first formed on 9/1/2019 and lasted for a total 11 days till 9/11/2019. We do, however, note that some cases may have the nearest detected event cluster(s) to be formed after the leakage event is reported with a maximum delayed time of 1 day, as demonstrated in the examples (see Table 3) for STN_B and STN_C in Zone-2, and STN_B in Zone-3. For all stations in Table 3, it is worth highlighting that the other leakage event clusters, which are not located temporally close to their respective reported events, may or may not represent hidden and unreported leakage events that require further field investigations.

4.2.2 Leakage localization

For each of the nearest event cluster detected to the reported events in Table 3, we proceed to estimate their dominant localization distances (D_{leak}) for every detected outlier within the event cluster by following our proposed mathematical procedures as summarized previously. Figures 8a-8c illustrates the resulting localization distances computed over the temporal size of the event cluster for a single leakage scenario from each of the zones, respectively:

- STN_A from Zone-1 for leakage event reported on 9/11/2019 (see Figure 8a)
- STN_A from Zone-2 for leakage event reported on 1/9/2020 (see Figure 8b)
- STN_C from Zone-3 for leakage event reported on 3/28/2020 (see Figure 8c)

In each of the Figures 8a-8c, several important pointers must be noted, namely: (1) the estimated D_{leak} values, as represented in the respective primary axis, over the temporal size of the event cluster are based upon harmonic mean computations in a rolling-forward temporal basis to emulate the near real-time context, (2) same harmonic mean computation principle is applied to compute the acoustic power over time in near real-time as represented in the corresponding secondary axis, and (3) the localization distance estimated for the final outlier of the event cluster, before the cluster breaks off, is then taken as the final average D_{leak} for the analysis.

By adopting the universal model configurations of 100-450Hz frequency range, N_{avg} = 12, 95% confidence interval for the upper-, lower- and vertical-bounds, $T_{thres} \approx 0.676$s, A_{thres} = 0.25, v_{air} = 1480m/s, and d_{width} = 50.0m, the final column of Table 3 summarizes the final average predicted D_{leak} value for each of the reported events. In summary, the following key observations can be made at this stage, namely:

- For most cases, the average error discrepancy between the reported and predicted pipeline distances is approximately 150m in absolute value, except for STN_B (Zone-1), STN_E and STN_F (Zone-2), and STN_A (Zone-3) where the bulk of their predicted distances are underestimated by more than 150m from the actual reported distances.

- Conservatively, the minimum detection criteria of $P_{ref} = 0.45$, $N_{outlier} = 1$, and $N_{cons} = 3$ provide sufficiently high confidence level that a pipeline leakage event is most likely taking place in the near proximity (≤ 600m) of the respective hydrophone station that is detecting and reporting the event cluster to the operator.

- As demonstrated in Figures 8a-8c, the normalized acoustic power values gradually increase over time, or at the very least, maintain a near-constant value above the minimum required power of 45% and above. This temporal observation can serve as an additional indication that a pipeline leakage event is taking place in the near proximity of the respective hydrophone station. Another common observation is that the power values may first rise to high value (e.g., to around 70-80%), followed by approaching a near-plateau power percentage value (> 45%) over the temporal size of detected event cluster.

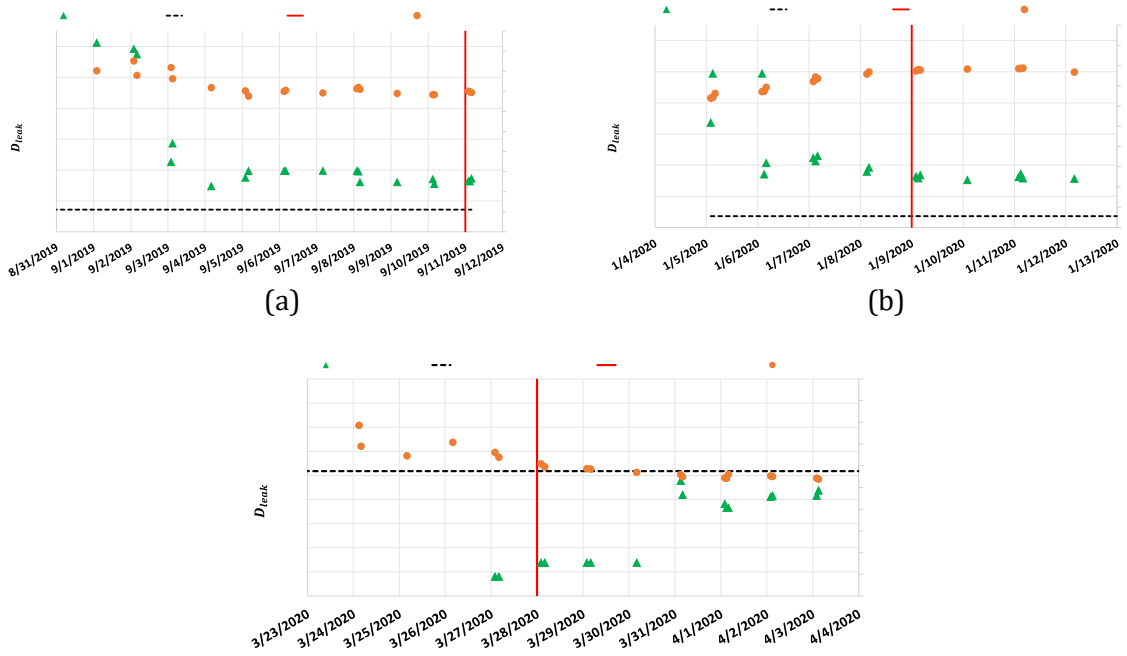


Figure 8. Estimated localization distances over temporal sizes of detected cluster located closest to respective reported leakage events: (a) STN_A from Zone-1 for leakage event reported on 9/11/2019; (b) STN_A from Zone-2 for leakage event reported on 1/9/2020; and (c) STN_C from Zone-3 for leakage event reported on 3/28/2020.

5 CONCLUSIONS

This paper develops a generalized acoustic data analysis methodology to perform near real-time leakage detection and localization in underground water distribution networks (WDNs). In collaboration with PUB, Singapore's National Water Agency, our proposed methodology comprises of systematic procedures for analysing acoustic data signals, namely: (1) data quality assessment; (2) features generations; (3) data pre-processing; (4) near real-time leakage detection and clustering; and (5) near real-time leakage localization. It is believed that the methodology can detect and localize leakage events in large WDNs having permanently installed acoustic sensors and has since been verified with 3 WDN zones in Singapore having 82 permanently installed hydrophone sensors across the networks. By emulating the near real-time context using historically available reported leakage events, our approach could successfully detect leakage events, as reported to within 600m or less from a neighbouring hydrophone station, with a maximum delayed time of 1 day in all 3 zones. The detected leakage event clusters are then further analysed to predict the likely locations of the leakage events from the nearest hydrophone stations, where the bulk of the events can be localized to within 150m error discrepancy, on average, with significant detected acoustic power of more than 45% over the temporal size of the event clusters.

6 ACKNOWLEDGEMENTS

This research is supported by the Singapore National Research Foundation under its Competitive Research Program (CRP) (Water) and administered by PUB (PUB-1804-0087), Singapore's National Water Agency.

7 REFERENCES

[1] E "2 2 p ' " p // p . / p - content/uploads/2020/12/Drinking-Water-2021.PDF, accessed July 2021, pp. 34–43, 2021, doi: 10.1061/9780784478851.

[2] . U. . p "M b W " W . 1, no. 1, p. 4, 2016, doi: 10.1186/s40713-016-0004-4.

[3] . M. . J p . W . . " oustic Filtering of the Pipe and Sensor in a Buried Plastic Water Pipe and its Effect on Leak Detection: An Experimental I " . 4 . 3. pp. 5595–5610, 2014, doi: 10.3390/s140305595.

[4] . M. " f an In- p W L " M Institute of Technology, 2010.

[5] . . K . K R. M M. . H b b " L W p U M p " J. p . Eng. Pract., vol. 3, no. 2, pp. 47–54, 2012, doi: 10.1061/(asce)ps.1949-1204.0000089.

[6] M. M. . "L p p b p p " I . . . 66 2 5 : 10.1063/1.4915721.

[7] . Z. . M J. W " b x p p " p p . . . 82 p. 108255, 2021, doi: 10.1016/j.apacoust.2021.108255.

[8] M. K Z. L M. . . E . H. " p I p p – Based Cross-Correlation Techniques along with Empirical Mode Decomposition for Water Pipeline Leakage Localization Utilizing Acousto- p " J. p . E . . . . 3 p. 4 2 27 2020, doi: 10.1061/(asce)ps.1949-1204.0000471.

[9] . H. M . . R. "I p L Iron and Copper Water- b p " J. p . E . . . 8 . 3 p. 5 7 2 7 doi: 10.1061/(asce)ps.1949-1204.0000257.

[10] R. . . . T J. " L W b p U p Autoencoder and H p " J. p . . E . . 34 . 2 p. 4 2 2 2 doi: 10.1061/(asce)cp.1943-5487.0000881.

[11] J. K . . M b . J. L . W . E " L b E b " IEEE Trans. Ind. Electron., vol. 65, no. 5, pp. 4279–4289, 2018.

[12] . . "L W b T –Frequency Convolutional " J. W R . . M . . 47 . 2 p. 4 2 2 2 : 10.1061/(asce)wr.1943-5452.0001317.

[13] . M J. . Z . M L. x M. . L b "L -Before-Break Main Failure W b p U W T : " J. Water Resour. Plan. Manag., vol. 146, no. 10, p. 05020020, 2020, doi: 10.1061/(asce)wr.1943-5452.0001266.

[14] . Z M. . L b M. L. p J. . . " p b prevention by permanent acoustic noise level monitoring in smart " U b W J. . 7 no. 9, pp. 827–837, 2020, doi: 10.1080/1573062X.2020.1828501.