Experimental investigation into techniques to predict leak shapes in water distribution systems using vibration measurements

Joseph D. Butterfield, Gregory Meyers, Viviana Meruane, Richard P. Collins and Stephen B. M. Beck

ABSTRACT

Water loss from leaking pipes represents a substantial loss of revenue as well as environmental and public health concerns. Leak location is normally identified by placing sensors either side of the leak and recording and analysing the leak noise. The leak noise contains information about the leak's characteristics, including its shape. Whilst a tool which non-invasively provides information about a leak's shape from the leak noise would be useful for water industry practitioners, no tool currently exists. This study evaluates the effect of various leak shapes on the vibration signal and presents a unique methodology for predicting the leak shape from the vibration signal. An innovative signal processing technique which utilises the machine learning method random forest classifiers is used in combination with a number of signal features in order to develop a leak shape prediction algorithm. The results demonstrate a robust methodology for predicting leak shape at several leak flow rates within several backfill types, providing a useful tool for water companies to assess leak repair based on leak shape.

Key words | pipeline, leakage, random forest, signal processing, water loss

Joseph D. Butterfield (corresponding author)
Stephen B. M. Beck
Department of Mechanical Engineering,
University of Sheffield,
Sheffield,
UK
E-mail: *jbutterfield1@sheffield.ac.uk*

Gregory Meyers
College of Engineering, Mathematics and Physical Sciences,
University of Exeter,
Exeter,
UK

Viviana Meruane
Department of Mechanical Engineering,
Universidad de Chile,
Santiago,
Chile

Richard P. Collins
Department of Civil and Structural Engineering,
University of Sheffield,
Sheffield,
UK

INTRODUCTION

Leakage from water distribution systems (WDS) results in a number of negative consequences, including revenue loss and environmental and public health concerns. A variety of leaks exists in pipelines, of different shapes, sizes and under different backfill types. Water loss represents a significant proportion of the distributed water, and therefore a substantial amount of research is focussed on developing new techniques in order to reduce water loss and locate the position of leaks. Traditionally, leakage levels are reduced through pressure management which is known to be a useful technique (Van Zyl & Cassa 2014), whereby the pressure within a zone (or 'district metered area') is optimised to reduce leakage and maintain a certain level of pressure at customer taps. Pressure management techniques are based on the orifice equation, whereby leak flow rate can be quantified by:

$$q = C_d A \sqrt{2gh} \tag{1}$$

where C_d is the discharge coefficient, g gravity acceleration, A hole area, h pressure head and q is the leak flow rate. Numerous studies have investigated the applicability of the orifice equation to leaks in WDS, resulting in the development of the power equation which is the preferred method

doi: 10.2166/hydro.2018.117

to model this relationship (Cassa & Van Zyl 2011):

$$q = ch^{\alpha} \tag{2}$$

where α is the leakage exponent and c a leakage coefficient. Evidently there is a strong relationship between pressure and leak area, and the response to pressure is also governed by the leak shape. However, within plastic pipe, this relationship becomes more complex due to pipe hysteresis (Ferrante 2011; Ferrante *et al.* 2011). It was reported by Almeida *et al.* (2014) that longitudinal cracks grow with pressure and time, whilst there is negligible growth of round holes in viscoelastic pipe. As the use of PE pipes is now much more common due to the assumed increased durability (GPSUK 2014), the phenomena of increased crack growth could result in increased leak flow rates and therefore greater water loss in WDS. Accurate quantification of leak flow rate can help to inform water companies and prioritise leak repair strategies. In the UK, water companies work towards a 'Sustainable Economic Level of Leakage' (SELL) which requires water companies to repair leaks providing this is cheaper than not fixing the leak (Ofwat 2002). Therefore, knowledge of the leak flow rate is vital in making operational decisions on whether or not to repair.

Although pressure management provides a useful technique for reducing leakage levels, the only way to completely remove water losses occurring due to the presence of leaks is by locating and repairing leaks. A common method to do this is through leak noise correlation (Puust *et al.* 2010). As water discharges from a leak, turbulence around the leak hole is created which transmits a signal in the form of vibration and acoustic waves, along the pipe wall and through the fluid (Papastefanou 2011). This is known as a leaks vibro-acoustic emission (VAE) signal. Sensors (usually accelerometers or hydrophones) are placed either side of the leak, recording and analysing the leak's VAE. The signals are then cross-correlated in order to identify the location of the leak. It has been noted that a number of factors influence a leak signal including the pipe material, backfill, flow rate and leak shape (Hunaidi & Chu 1999; Muggleton & Brennan 2004; Pal 2008; Butterfield *et al.* 2017a, 2017b).

Leaks commonly studied in pipelines in terms of VAE signals are normally one of three types: round holes (Wassaf *et al.* 1985; Pal 2008; Butterfield *et al.* 2016a, 2016b, 2017a, 2017b); artificial leaks from fire hydrants (Almeida *et al.* 2014; Gao *et al.* 2015); and slits (Wassaf *et al.* 1985; Pal 2008). However, a wider variety of leak shapes can occur in WDS, including pin holes, joint leaks, circumferential slits and longitudinal slits (UKWIR 2008), amongst others. Although there is a large variety of leak shapes in pipelines, which is likely to influence the accuracy of leak noise correlation, there has been limited study investigating the effects of the leak shape on the VAE signal. Brunner & Barbezat (2007) simulated round holes of different diameters and subsequently found differences between the power spectra of different leak sizes. Wassaf *et al.* (1985) compared the acoustic emission responses of circular holes and rectangular slits, showing that the shape influenced the signals frequency spectrum. The use of standpipes to create artificial leaks (Hunaidi & Chu 1999; Ferrante *et al.* 2011; Butterfield *et al.* 2017a, 2017b) is also common in the literature, but these are not representative of real leaks in WDS. Pal (2008) compared the signals of artificial leaks created by fire hydrants, with joint leaks made by loosening the nuts on a flange plate and split leaks. They found that the frequency spectrum of a leak was strongly governed by its shape and size. However, the leak flow rates were not controlled in this study and therefore were different for each leak shape. As the leak flow rate has been shown to have such a strong influence on the leak signal (Butterfield *et al.* 2017a, 2017b), any assessment of a leak shape requires a good experimental methodology which controls the leak flow rate between shapes and thus isolating the effect of leak shape on the leak signal.

Although the aforementioned studies provide a useful insight into the effect of leak shapes on the VAE signal, the leak shapes studied are uncommon on plastic pipes in real WDS (Water Services Association of Australia 2012) and therefore have limited representation of real leaks in real WDS. The majority of leaks in plastic pipes actually occur due to leaky joints (Water Services Association of Australia 2012; Tayefi 2014), which is typically due to contamination of the joint when the pipe is being installed. This is especially true with electrofusion and butt fusion joints (Tayefi 2014). However, there have been no studies investigating the VAE signal produced by leaky electrofusion joints in laboratory conditions. This highlights a

significant research gap as leaks representative of 'real leaks' in plastic pipe are seldom compared and there has been little comparison of leak shapes.

A study by Sun *et al.* (2016) analysed the effect of varying leak areas for round holes in gas pipes. Their study went one step further showing that the area of the hole could be predicted from the VAE signal in combination with Support Vector Machine (SVM). Although they only investigated round holes, the classification of a leak shape is a useful tool. The ability to distinguish between leak shapes would allow prioritisation of leak repair by fixing those leaks more likely to grow. Despite the potential shown in gas pipes by Sun *et al.* (2016), a method to identify leak shapes with VAE in WDS does not currently exist. In this paper a novel medium density polyethylene (MDPE) pipe rig was developed and used to simulate various leak shapes at different sizes and flow rates. A comparison of the leak signals is made and then a methodology for predicting the leak shape from only the VAE signal in combination with a random forest machine learning model is presented. This paper therefore presents a new method for predicting leak shape in WDS from vibration signals.

LEAK SHAPE PREDICTION

Due to the wide variety of variables influencing a leak signal, a machine learning approach to leak shape prediction may provide the most robust results. Any model relies on the derivation of a number of signal features which describe the leak and therefore enable the prediction of leak shape. Leak VAE signals have been shown to differ in both time and frequency domains (Ahadi & Bakhtiar 2010; Butterfield *et al.* 2016a, 2016b) and therefore time-frequency based features could provide useful features. Feature extraction methods such as the wavelet transform and empirical mode decomposition (EMD) provide good time-frequency resolution. EMD derives the intrinsic mode functions (IMFs) through the following procedure:

1. Locate signal extrema $x'(t)$.
2. Calculate upper and low envelope connecting the minima (cf. minima) and maxima (cf. maxima), $e_{min}(t)$ (cf. $(e_{max}(t))$ by interpolating (spline interpolation).
3. Calculate mean between lower and upper envelopes, $m(t) = (e_{min}(t) + e_{max}(t))/2$.
4. Subtract mean obtaining modulated oscillation [24], $d(t) = x'(t) - m(t)$.
5. Apply stopping criteria (Mandic *et al.* 2013). If $d(t)$ satisfies stopping criteria, let $d(t)$ become IMF_m.
6. Subtract the new IMF from signal $(x'(t))$, so $x'(t) := x'(t) - IMF_m$.
7. Sift until IMF calculated in step 5 becomes a monotonic function.

Sun *et al.* (2016) found that taking the root mean square (RMS) of individual IMFs was a useful feature in classifying leak diameter, however EMD has also been demonstrated to be limited by a mode mixing problem whereby the physical meaning of signal can be lost (Huang *et al.* 1998). To overcome this problem, the ensemble empirical mode decomposition (EEMD) was developed by Wu & Huang (2005) whereby Gaussian white noise is added to the signal. The EEMD is given by:

$$\text{Ensemble: } \{S_n(t)\}_{n=1}^{N} = x(t) + \{w_n(t)\}_{n=1}^{N}, \tag{3}$$

where $\{w_n(t)\}_{n=1}^{N}\ N(0, \sigma)$ indicates Gaussian noise and $x(t)$ represents the leak signal.

The RMS of the raw time-domain signal was described by Butterfield *et al.* (2017a, 2017b) to correlate well with increasing leak flow rate. Shannon entropy following local mean decomposition was used by Sheng *et al.* (2016) to predict bearing condition and Sun *et al.* (2016) used Shannon entropy of individual IMFs in order to predict leak aperture in gas pipes. The maximum and mean dB of a signal's power spectral density was used by Chen *et al.* (2007) to describe leak flow rate in gas pipes. Prime & Shevitz (1996) found that the fundamental frequency varied due to cracks in beams. Spectral shape methods such as kurtosis, skewness (Kakur & Jurecka n.d.) and Crest factor (Pachoud *et al.* 2009) as well as signal descriptors such as the standard deviation and signal power (Picone 1993) have been used in speech and audio recognition/detection type problems and have shown to be useful features. A similar set of features was used by Butterfield *et al.* (2017a, 2017b) for the quantification of leak flow rates in plastic pipe.

Crucial to the task of predicting leak shape is use of pattern recognition algorithms. Random forest (models) have been shown encouraging results in identifying patterns in speech (Su *et al.* 2007) and a similar approach may recognise patterns in leak VAE signals. Random forest (RF) models are a machine learning method that can be considered as an ensemble of many decision trees, where each decision tree is trained to optimally split the data into separate classes (Breiman 2001). Each decision tree is provided with a random subset of the data for training and identifies the best split between classes based on a small subset of features. Therefore each individual decision tree alone is a weak classifier, however good performance, scalability and generalisation can be obtained by combining all decision trees in the ensemble. Class prediction on unseen test data is then obtained from a majority vote from the ensemble. A probability can also be calculated of how certain the model is in the prediction.

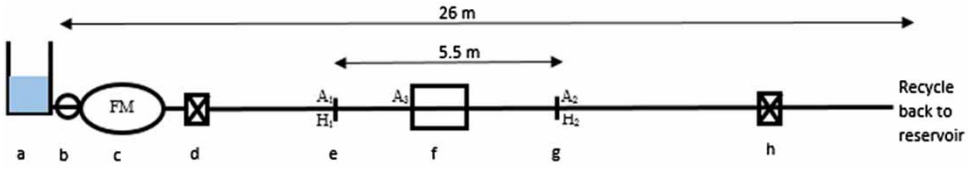
METHODOLOGY

Experimental setup

The state-of-the-art LiVE (leaks in viscoelastics) pipe rig at the University of Sheffield, UK is used in this study. The rig consists of a 26 m long MDPE, 63 mm diameter pipe loop. A schematic of the pipe rig is shown in Figure 1. Water is supplied to the pipe rig using a 3.5 kW variable speed pump (Wilo, Burton-on-Trent, UK) from an upstream reservoir (0.95 m^3 by volume). Water passes a magnetic flow meter (Flow Systems 91DE) recording system flow rate. System pressure is measured with two pressure sensors (Gems Plainville 2200) located upstream and downstream of the leak, recording at a sample rate of 2,000 Hz.

A 5.5 m long 'test section' is located in the middle of the pipe rig (indicated between points e and g in Figure 1). This section of pipe is removable at two flange plates and was used in order to create leaks of different shapes and sizes. Round holes measuring various diameters and longitudinal slits were drilled using standard drill bits. Two leaky electrofusion joints of different sizes were created by excavating a small void from the pipe wall before welding two pipe sections together. This ensured that a gap measuring the size of the void was present and water could discharge through the remaining hole. A schematic is shown in Figure 2 in order to illustrate this process. The leak shapes were all sized in order to have equivalent leak areas of ~10, ~16, ~24 and ~32 mm^2. Details of the exact dimensions of the leak shapes investigated are shown in Table 1 along with the total number of simulations per leak flow rate. Ranging in area size, a total of four round holes, four longitudinal slits and two leaky electrofusion joints were tested. Each leak shape was drilled into a separate length of straight pipe (measuring 5.5 m in length) and inserted into the 'test section'. Photographs of the leak shapes are shown in Figure 3.

The leaks discharged into a $0.5 \times 0.5 \times 0.5$ m cubic box. In order to simulate the influence of backfill type on the leak signal, the backfill type was changed for each test. The cubic box was filled with 5–12 mm diameter pea gravel backfill in accordance with British Standards for backfill of plastic pipe (BSI 1973) and therefore represents a standard external porous media. The alternative backfill



a. Reservoir
b. Pump
c. Flow meter
d. Upstream butterfly valve
e. Upstream flange plate
f. Test section box (0.5 x 0.5 x 0.5 m)
g. Downstream flange plate
h. Downstream gate valve
A_1. Accelerometer 1
A_2. Accelerometer 2
A_3. Accelerometer 3
H1. Hydrophone 1
H2. Hydrophone 2

Figure 1 | Schematic of the pipe rig (not to scale). Adapted from Butterfield *et al.* (2016a, 2016b).

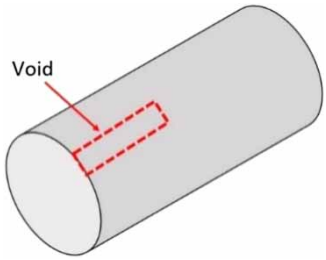


Figure 2 | Demonstration of the void created before welding the electrofusion joint.

Table 1 | Leak shape and sizes

Area (mm^2)	Flow (L/min)	# of Slit samples	# of Round samples	# of Electro samples	Total samples[a]
9.62–10	39	20	20	0	40
	44	20	20	0	40
	47	20	20	0	40
	49	20	20	0	40
	56	0	0	0	0
15.9–16	39	20	20	20	60
	44	20	20	20	60
	47	20	20	20	60
	49	20	20	20	60
	56	20	20	20	60
23.76–24	39	20	20	0	40
	44	20	20	0	40
	47	20	20	0	40
	49	20	20	0	40
	56	20	20	0	40
32–33.18	39	20	20	20	60
	44	20	20	20	60
	47	20	20	20	60
	49	20	20	20	60
	56	20	20	20	60
Totals		380	380	200	960

[a]The number of samples for all leak shapes.

types were geotextile fabric and completely submerging the pipe in water. The geotextile fabric was also used by Fox (2016) and represents a constrained external porous media. Photographs of the external media types are shown in Figure 4. Five different leak flow rates (approximately 39–40, 44–45, 47–48, 49–51 and 56–57 L/min were used in the simulations. These were kept the same across all leak shapes and area sizes by controlling system pressure with the downstream gate valve. Therefore, each leak shape is assessed on three backfill types at five different leak flow rates. The pump was run continuously whilst acquiring data for each simulation, but was turned off when replacing the test section and varying backfill. The wave speed in the pipe rig is estimated to be 347 m/s using theoretical calculations (Almeida *et al.* 2014).

Signal processing

The leak's VAE signal was recorded at 2,500 Hz using an accelerometer (PCB Piezotronics 393B12, sensitivity 10 V/g) placed approximately 30 cm away from the leak. This sensor was powered by a current source power unit (Dytran Instruments type 4102C). Signals were digitised using a data acquisition unit (National Instruments cDAQ) and imported into Matlab. Signals were then pre-processed with a 4th Order Butterworth bandpass filter set at 10–1,000 Hz. Each signal sample was measured with the accelerometer for 5 seconds.

Due to physical phenomena within the pipe rig, 20 samples from each simulated leak shape, area size and leak flow rate were taken to examine the variation between samples. It was found that there was little variation between samples, suggesting repeatable results. Table 1 shows the aggregate of the number of samples and simulations carried out over all the flow rates. The smallest of the area sizes for each leak shape has fewer simulations as they were too small for the highest flow rate of 56–57 L/min to be achieved. Table 2 lists the features extracted from each leak VAE signal used in this paper.

The RF in this paper consists of 1,000 decision trees and the entropy splitting criterion was used to measure the quality of a split. The number of features considered when looking for the best decision tree split and the maximum depth of each decision tree was determined by hyperparameter tuning using 5-fold cross-validation on the training data.

Another five-fold cross-validation was used for splitting the training and test datasets, also known as nested cross-validation. In this outer five-fold cross-validation data are split into five equally sized sections and the model is trained on four sections (i.e. 80% of the data) and tested on the

Figure 3 | Photographs of leaks: (a) round hole; (b) longitudinal slit; (c) electrofusion joint.

Figure 4 | Photographs of leaks in different backfill: (a) gravel media; (b) submerged; (c) geotextile fabric.

Table 2 | Derived features from the VAE signal

Feature no.	Name
1–6	RMS of IMFs1-6
7–12	Shannon entropy of IMFs 1–6
13	Shannon entropy of whole signal
14	RMS of whole signal
15	Mean dB of power spectral density
16	Maximum dB of power spectral density
17	Minimum dB of power spectral density
18	Standard deviation
19	Signal power
20	Fundamental frequency
21	Spectral flux
22	Kurtosis
23	Skewness
24	Crest factor

remainder section (i.e. 20%). Four additional identical models are each independently trained using a sum of four sections worth of data but each model is tested using a different remaining final section. This way all the data were used in training and testing but no individual model is trained on its testing data. The average test accuracy of all five model results in an accuracy score that is almost completely unbiased of how the data were split up into training and testing sets (Varma & Simon 2006). A process flow diagram of the whole experimental, signal processing and machine learning programme is shown in Figure 5.

RESULTS

Characteristics of leaks with different shapes

VAE signals from the three different leak shapes of equivalent leak area are plotted in Figure 6 at ~40 L/min in the frequency domain as a representation of the ratio of leak: no-leak in order to fully demonstrate the contribution of the leak shape to the leak signal. The contribution of the leak noise to the received signal is at frequencies >63 Hz for all leak shapes. The leak signal has a wide spectral

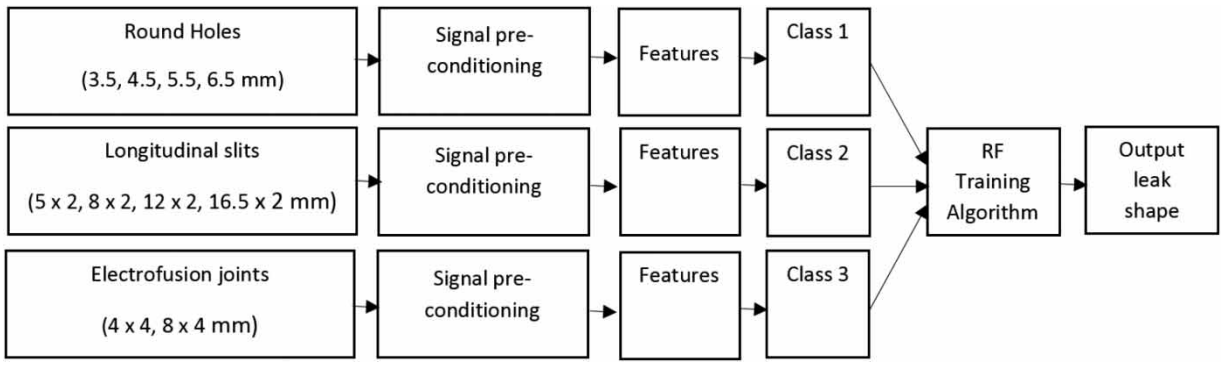


Figure 5 | Process flow diagram.

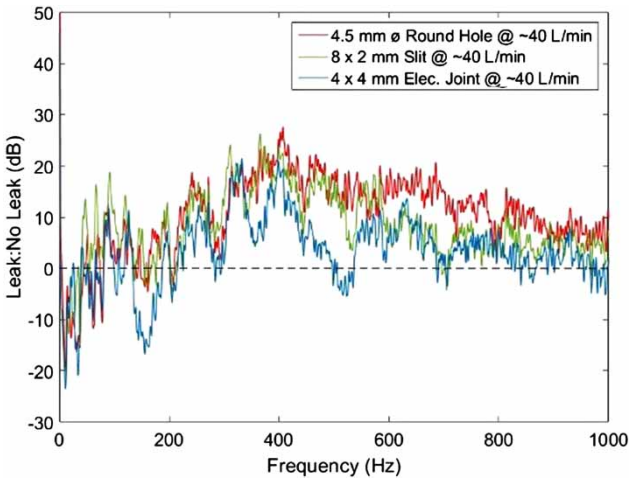


Figure 6 | Ratio of leak:no-leak for comparing different leak shapes at ~40 L/min leak flow rate.

range which differs depending on leak shape. The highest amplitude signals appear between 250 and 570 Hz for all leak shapes. Between frequencies 200 and 300 Hz, leak signals are very similar for all leak shapes. The round holes tended to have a slightly wider spectral range compared to the other two leak shapes. The electrofusion joint had the lowest amplitude signals, whereas the round holes were consistently the highest amplitude at frequencies >250 Hz. The round hole and longitudinal slit observed a steady decline in amplitude at frequencies >570 Hz, although this was more rapid for the longitudinal slit. The electrofusion joint, however, decreases amplitude rapidly at frequencies >430 Hz. The electrofusion joint and longitudinal slit actually became markedly similar at frequencies of 560 Hz.

Feature extraction of leak signals

The signal processing method involved the use of 24 different features which were extracted from the signal in time and frequency domains. All leak shapes and sizes were decomposed via EEMD generating individual IMFs. These IMFs were transposed in to the frequency domain via the Fourier transform for comparative purposes and the frequency spectrums of the first six IMFs are presented in Figure 7.

All IMFs represent signals of low frequency (<800 Hz). The highest frequency signals within this range are located within the first IMF (IMF1), but this was distinctly low amplitude. Frequency decreases as IMF number increases. The comparison between leak shapes reveals differing frequency distributions across different IMFs depending on leak shape. IMF1 is mainly dominated by signals from the round hole which has the widest spectral band of all the IMFs, the electrofusion joint is largely within IMFs 2, 3 and 4 whilst the longitudinal slit is dominant in IMF5 and 6. The electrofusion joint appears to have power in IMF2, 3 and 4 whilst the longitudinal slit has more power in IMF5. It is therefore possible that an analysis of the IMFs could provide information on the leak shape.

Classifying leak shape

Shape classification

Only the 24 features derived from the accelerometer signal were used as inputs to each model, e.g. the models were

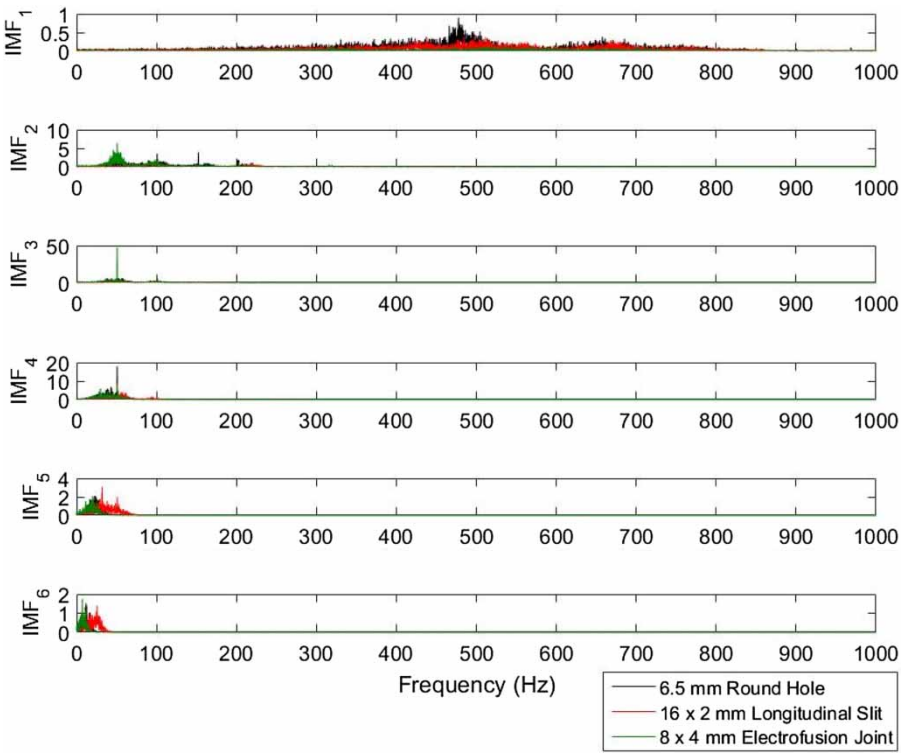


Figure 7 | Comparison of leak shapes at individual IMFs following EEMD. All leak flow rates are set between 47 and 49 L/min and equivalent hole area is 32–33.18 mm^2.

not told the leak shape or the leak flow rate. All leak shapes were initially divided up by their leak area, creating five separate datasets: 10, 16, 24 and 33 mm^2; and All (all leak areas used and the model not told the leak area size). As the leak shape corresponds to a matching leak area of another shape, the effect of leak shape is isolated. As electrofusion joint data were only available for datasets of leak areas 16 and 33 mm^2, the datasets 10 and 24 mm^2 contained fewer input data.

The performances of the models in classifying leak shape by individual areas is shown in the form of confusion matrixes in Figure 8. The confusion matrices are further divided by leak area size, the 10 mm^2 dataset (Figure 8(a)), 16 mm^2 dataset (Figure 8(b)), 24 mm^2 dataset (Figure 8(c)), 33 mm^2 dataset (Figure 8(d)) and All leak area sizes (Figure 8(e)). The model was able to classify leak shape for all areas at all leak flow rates within all backfill types to >81%. Generally, it appeared that classification accuracy was notably higher for round holes compared to the other leak shapes. In the case of the 24 mm^2, the model correctly classified 99% accuracy. An investigation into the 'All' dataset (Figure 8(e)) suggests that the round holes also achieved the highest prediction accuracy at 90%, followed by the electrofusion joint at 81%. The worst performance in this dataset was the longitudinal slit with a classification accuracy of 75%.

Figure 8(e) demonstrates the average classification accuracy for each leak area, each leak shape and each leak flow rate when using the 'All' dataset. Note, this is not the averaged-output of the model in Figure 8, it is the individual subset results for the model using the 'All' dataset. An investigation into the performance of the model for different leak areas within the 'All' dataset shows increased average classification accuracy with the 24 mm^2 dataset (98%) (Table 3). However, this may be due to a smaller dataset as electrofusion joints are not included. Despite the fact that there were five different leak flow rates studied, the model was able to classify the leak shape independent of leak flow rate at >82% accuracy for all leak shapes at all leak areas.

Figure 9 demonstrates the breakdown of the performance of the RF model using the 'All' dataset for each leak shape by leak area (Figure 9(a)), leak flow rate

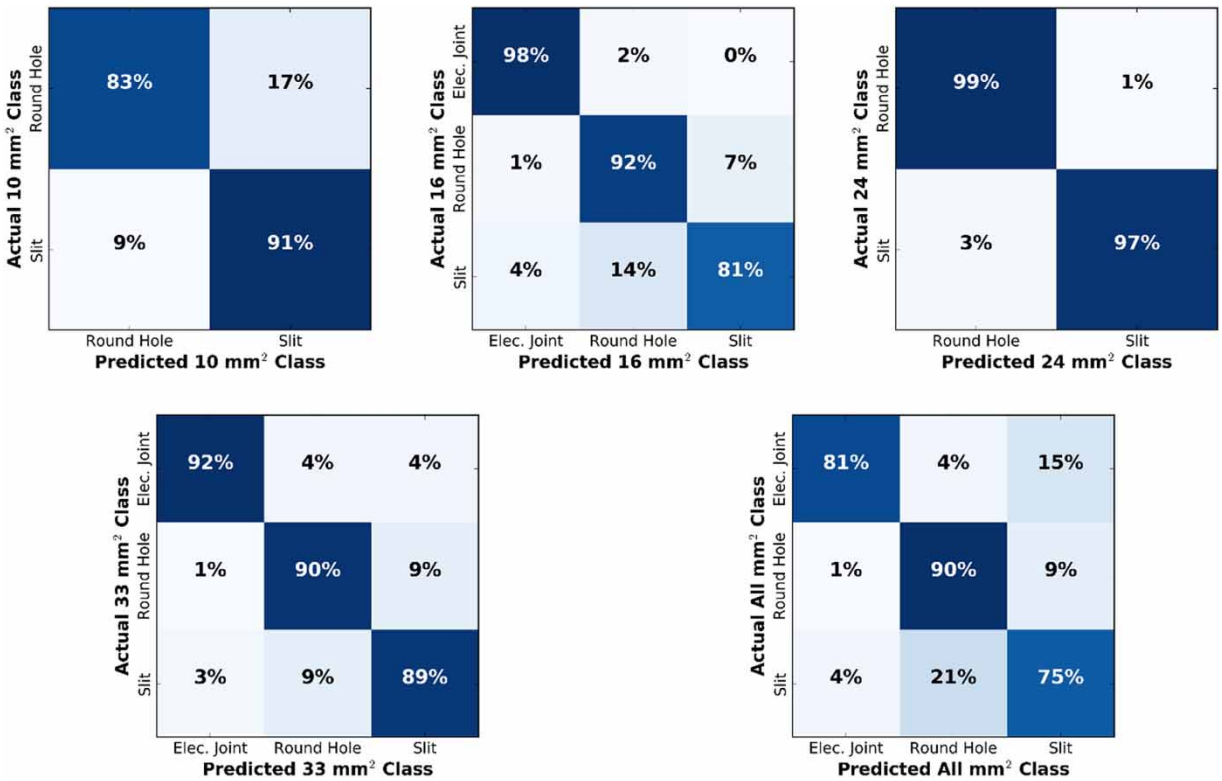


Figure 8 | Model accuracy for each hole shape by leak area.

Table 3 | Classification accuracies of the 'All' dataset

Dataset (mm^2)	Training classification accuracy (%)	Testing classification accuracy (%)
10	93	87
16	94	90
24	100	98
33	97	90
All	91	82

(Figure 9(b)) and backfill type (Figure 9(c)). For all leak areas studied, shape classification accuracy is greatest for round holes (>90% accuracy) (Figure 9(a)). The model performs well at 10 and 24 mm^2 (>85%), but again this may be due to the fact that these leak areas do not include the electrofusion joint data and therefore there is a smaller dataset. The breakdown of individual leak flow rates for the 'All' dataset shows that there is no observable trend between leak flow rate and shape prediction accuracy (Figure 9(b)). However, the round holes tended to have greater consistency in prediction accuracy at all flow rates. The electrofusion joint performed comparably to the round hole at the lower flow rates, whilst accuracy dropped to <68% at the higher leak flow rates. Generally, prediction of slits shape was similar at all leak flow rates, between 68% and 78%. The breakdown of the performance of the models within individual backfill types shows that at the lowest leak flow rates, classification of leak shape performed better on the gravel media (Figure 9(c)). However, prediction accuracy was improved under geotextile fabric at the highest leak flow rates. The submerged backfill tended to perform worst at the mid to high range leak flow rates.

Feature importance

Figure 10 ranks feature importance of the RF model by ranking the use of the 24 different features input into the model. This is broken down by the 'All' dataset into individual leak areas and all leak areas. It was found that the most important feature to the model was the RMS of IMF1. This was

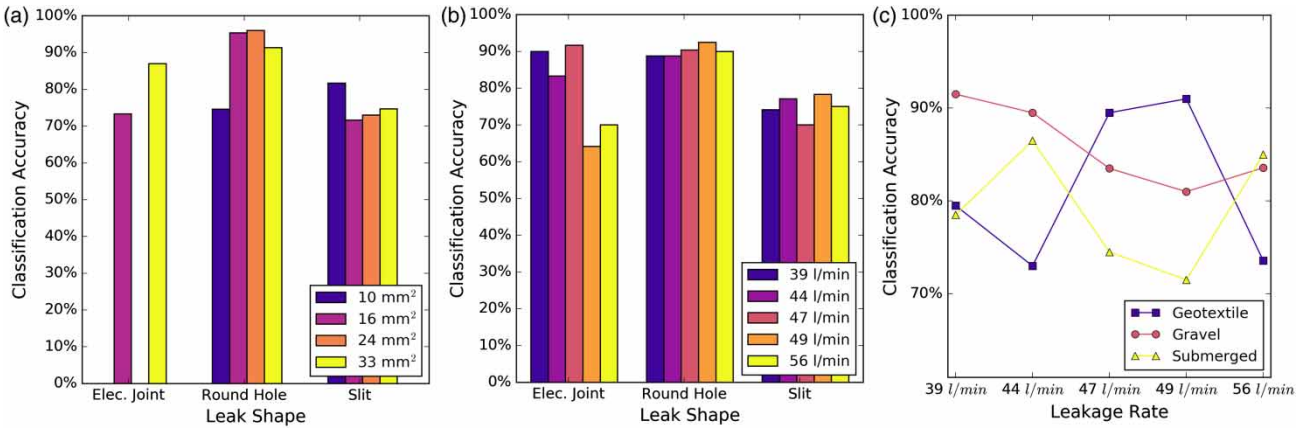


Figure 9 | Classification accuracy by different subsets: (a) leak shape by shape area; (b) leak shape by leakage rate; (c) leakage rate by media type.

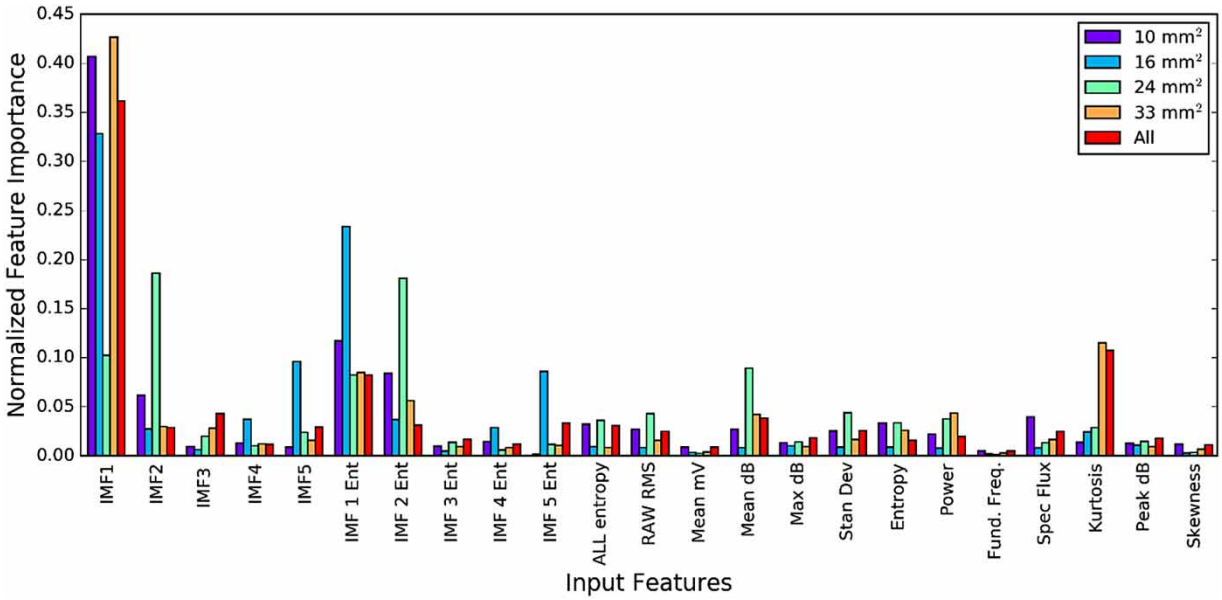


Figure 10 | Feature importance when classifying leak shape.

true no matter what the leak area and the leaks shape. The model also found other features useful, however the degree to which the model found these features important depended on whether the model was predicting based on individual leak areas of the 'All' dataset.

Effect of different backfill types on classification accuracy

In real WDS, a variety of backfill types exist and the backfill has been shown to influence the leak signal. The effect of backfill type on model performance was evaluated by training and testing on individual backfill types rather than all backfill types. The results for this are shown in Table 4. Evidently, backfill type has a large impact on the performance of leak shape prediction. Overall, training on only one type of backfill and testing the model on a separate backfill type had a largely negative impact on model performance. The worst performance appeared to be training the model on gravel but then testing on submerged data. Therefore, either backfill type should be known or the model needs to be trained and tested on more backfill types.

Table 4 | Model accuracy on non-trained media

Trained on	Tested on	Testing accuracy (%)
Gravel only	Geotextile	41
	Submerged	21
Geotextile only	Gravel	36
	Submerged	32
Submerged only	Gravel	25
	Geotextile	46

DISCUSSION

Influence of leak shape on leak signal characteristics

Leak shape was found to influence leak signal amplitude and frequency when the leak flow rate was isolated by controlling system pressure (Figure 6). Although there has been limited investigation in the literature, this study is coherent with existing literature study that the leak shape influences the leak signal (Pal 2008). However, across the whole shape spectrum (Figure 6), it is difficult to separate leak shapes based on frequency alone although this may be possible if just comparing electrofusion joints and round holes. Differences in the spectrum of leak shape were also identified when the signal was divided into different frequency bands using EEMD, where the power of the signal in each IMF differed depending on each leak shape (Figure 7). This appears to suggest that leak signals differ in both time and frequency domains, also shown by previous authors (Ahadi & Bakhtiar 2010; Butterfield *et al.* 2016a, 2016b). Differing signal spectra due to changes in leak shape may be due to varying jet angles (Ferrante *et al.* 2013) as the leak discharges the hole. In turn, this will likely create varying turbulence regimes around the leak hole specific for that leak shape, where the signal is created (Papastefanou 2011). Therefore, this study has experimentally determined that leak shape is another key variable in determining leakage behaviour in addition to leak flow rate and leak area (Cassa & Van Zyl 2011; Ferrante 2012) and these results can better inform the design of leak noise correlators.

Model performance and feature importance

The model presented herein demonstrates that it is possible to predict leak shape to high classification accuracy at all the leak flow rates and within all the backfill types studied (>80%). The use of this model provides practitioners with a tool to predict leak shape. Due to the fact that certain leak shapes have time- and pressure-dependent growth (Ferrante 2012; Fox 2016), knowledge of the leak shape will allow for prioritisation of leak repair. Moreover, the tool provides an opportunity for water companies to collect more data about the shapes of leaks, and thus this information can be linked to further parameters (such as pipe failure mode).

The RMS of IMF1 was found to be the most important feature when classifying leak shape (Figure 10). The Fourier spectrum of this feature demonstrates that the amplitude within this IMF is dependent on leak shape (Figure 7). Further investigation into this feature is demonstrated in Figure 11,

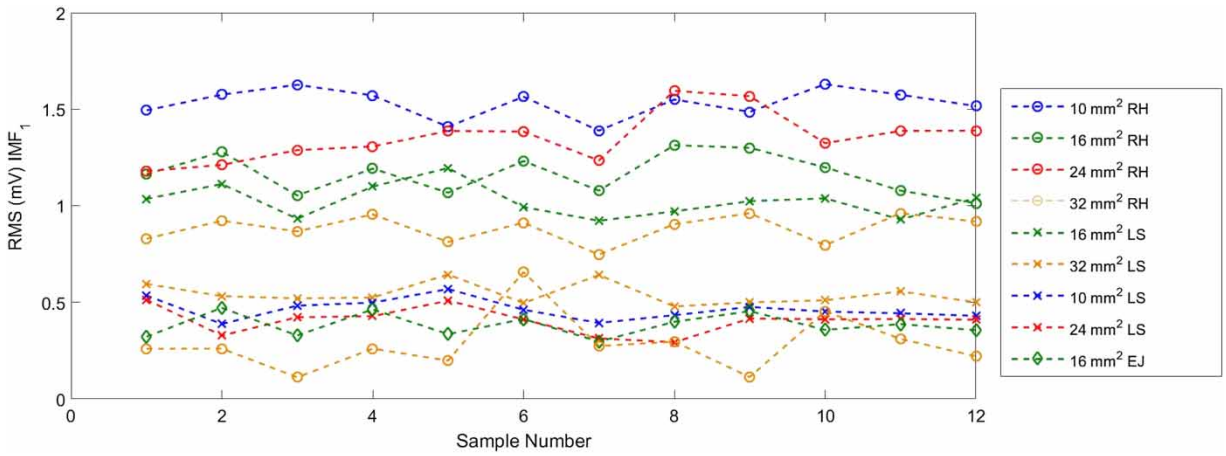


Figure 11 | RMS of IMF1 for all leak shapes of different leak areas at 44.7 L/min. RH = round hole; LS = longitudinal slit; and EJ = electrofusion joint.

showing the individual RMS of IMF1 for each leak shape and leak area. For all leak shapes, it appears generally possible to distinguish between all shapes (of all leak area) individually using this feature. However, in some cases it becomes difficult due to overlap between leak shapes, especially for the electrofusion joint at 32 mm^2. This highlights the necessity for a machine learning based tool to provide the best separation between signal features. Leak signal RMS has previously been shown to correlate well with an increase leak flow rate (Chen *et al.* 2007; Kaewwaewnoi *et al.* 2007; Papastefanou 2011) and therefore describes information about the leak signal. It was also found by Sun *et al.* (2016) that the RMS of each IMF could provide a good descriptor of the leak area. It is therefore logical that the IMF most related to the leak signal and the RMS of this would provide a good method of classifying the leak signal. However, this feature represents the higher frequency content (350–650 Hz) and when measuring further away from the leak these frequency bands would normally be attenuated due to the pipe acting like a low pass filter (de Almeida *et al.* 2015). Therefore, this feature will become less effective when moving sensors further away from the leak.

Influence of backfill type

The effect of pipe backfill was explored by altering the parameters of the model, training and test on differing combinations of backfill type (Table 4). Backfill type was found to strongly influence the accuracy of the RF model, most likely because the leak signal is strongly influenced by the surrounding backfill (Muggleton & Brennan 2004; Butterfield *et al.* 2016a, 2016b). Further investigation into the effect of backfill on the leak signal is shown in Figure 12 when the leak flow rate is consistent between backfill types (~47 L/min is shown). Similar frequency components between ~250 and 570 Hz were found for all backfill types. After 570 Hz, there was a drop in signal amplitude for all backfill types and this is in agreement with the results shown in Figure 12. However, the gravel backfill remained at a higher amplitude at frequencies >570 Hz, followed by geotextile and then submerged backfill types. Both the gravel and geotextile fabric are similar in that they are a constrained media type (Fox 2016), which will have an impact on the water jet as it leaves the leak hole. However, as is not possible to achieve good prediction results when training on gravel but testing on geotextile fabric, these results suggest that these media types play a differing role in impacting the generation or transmission of the leak noise. The submerged pipe had the lowest amplitude signals (Figure 12) but only at frequencies >570 Hz. A similar effect of backfill was noted by Muggleton & Brennan (2004) who found smaller signal attenuation in the submerged pipe compared to a pipe in soil, however there was a disappearance of the leak signal in the submerged pipe. These results demonstrate that the model is more robust when trained on more backfill types, and future work should include a wider range of conditions such as soil saturation, fluidisation and further types which have all been shown to influence leak-media hydraulics (van Zyl *et al.* 2013; Fox 2016). Moreover, the impact of these aforementioned variables is still a major research gap in terms of leak VAE signals.

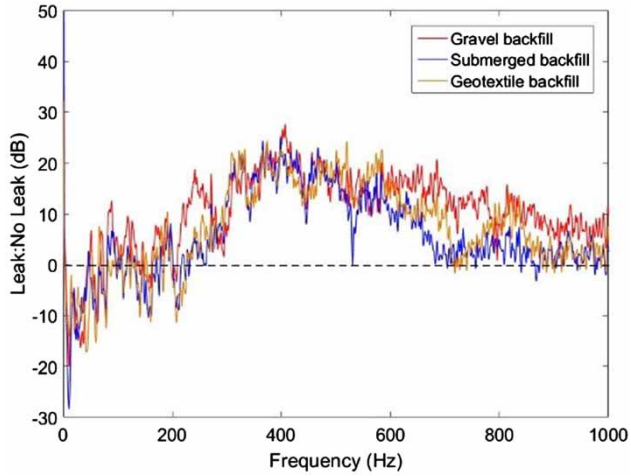


Figure 12 | Leak spectrum of a 12 × 2 mm longitudinal slit at 47 L/min leak flow rate under three media types compared to a no-leak scenario.

Study limitations

Whilst the tool provided herein provides a tool for water companies to prioritise leak repair, it is not without limitations. A key weakness of this study is the fact that leak shape prediction was undertaken approximately close to the leak. In real-world conditions, it is unlikely that any measurements would take place within such close proximity to the leak. In fact, accelerometers/hydrophones are normally placed on or in nearby fittings, such as valves and hydrants at some

distance away. The value of the model has not been tested at further distances, and this remains a key element of future work. However, a number of developments exist in other areas such as pipeline robotics (Chatzigeorgiou *et al.* 2014), which will allow a sensor to travel to a position next to a leak and the system developed here can possibly be integrated into these tools. This study also only addresses a limited number of leak shapes, flow rates, sizes and shapes under only three different backfill types. In real systems, the variety of leaks under varying conditions is huge, and therefore the validity of this system to real world leaks is not known. However, the study has shown that it is possible to differentiate between the leak shapes studied, independent of leak flow rate, leak area and backfill type.

CONCLUSIONS

The research presented herein has demonstrated that the leak's VAE signal contains enough information within it to predict the leak shape. A unique experimental investigation used high quality experimental data from various leak shapes of several leak areas at five leak flow rates. Leak shape was found to be a significant factor influencing the leak signal. Twenty-four features were derived from the leak signal and in combination with a random forest model it was possible to predict leak shape to a relatively high accuracy. It was also found that the external backfill had a strong impact on the classification accuracy, but training on all backfill types provided a more robust tool with higher classification accuracy. While the variety of leaks under varying conditions is huge, and therefore the validity of this system to real world leaks is not known, the proposed technique provided in this paper demonstrates that it is possible to predict the leak shape independent of leak area, leak flow rate and backfill type. Therefore, this investigative study is the first to demonstrate that there is enough information within a leak signal in order to predict the leak shape.

ACKNOWLEDGEMENTS

The authors would like to acknowledge and thank Northumbrian Water, Severn Trent Water, Thames Water Utilities, Scottish Water and the EPSRC under grant number EP/G037094/1 for their valued contribution to this research.

REFERENCES

Ahadi, M. & Bakhtiar, M. S. 2010 Leak detection in water-filled plastic pipes through the application of tuned wavelet transforms to acoustic emission signals. *Appl. Acoust.* **71**, 634–639.

Almeida, F., Brennan, M., Joseph, P., Whitfield, S., Dray, S. & Paschoalini, A. 2014 On the acoustic filtering of the pipe and sensor in a buried plastic water pipe and its effect on leak detection: an experimental investigation. *Sensors* **14**, 5595–5610.

Breiman, L. 2001 Random forests. *Mach. Learn.* **45**, 5–32.

Brunner, A. J. & Barbezat, M. 2007 Acoustic emission leak testing of pipes for pressurized gas using active fiber composite elements as sensors. *J. Acoust. Emission* **25**, 42–50.

BSI 1973 *CP 312-1: Code of Practice for Plastic Pipework (Thermoplastics Material) – Part 1: General Principles and Choice of Material*. British Standards Institution, UK.

Butterfield, J. D., Collins, R. P. & Beck, S. B. M. 2016a Feature extraction of leak signals in plastic water distribution pipes using the wavelet transform. In: *Proc. ASME 2015 Int. Mech. Eng. Congr. Expo.*, Houston, Texas, pp. 1–8.

Butterfield, J. D., Collins, R. P., Krynkin, A. & Beck, S. B. M. 2016b Experimental investigation into the influence of backfill types on the vibro-acoustic characteristics of leaks in MDPE pipe. *Procedia Eng.* **186**, 311–318.

Butterfield, J. D., Krynkin, A., Collins, R. P. & Beck, S. B. M. 2017a Experimental investigation into vibro-acoustic emission signal processing techniques to quantify leak flow rate in plastic water distribution pipes. *Appl. Acoust.* **119**, 146–155.

Butterfield, J. D., Meruane, V., Collins, R. P., Meyers, G. & Beck, S. B. M. 2017b Prediction of leak flow rate in plastic water distribution pipes using vibro-acoustic measurements. *Struct. Health Monit.* **18**, 1–12.

Cassa, A. M. & van Zyl, J. E. 2011 Predicting the head-area slopes and leakage exponents of cracks in pipes. In: *Proceedings of the Urban Water Management: Challenges and Opportunities (CCWI)*, 5–7 September. University of Exeter, Exeter, pp. 485–491.

Chatzigeorgiou, Y., Wu, D., Youcef-Toumi, R. & Ben-Mansour, K. 2014 MIT Leak Detector: An in-pipe leak detection robot. In: *Robot. Autom. (ICRA), 2014 IEEE Int. Conf. IEEE*, Hong Kong.

Chen, P., Chua, P. S. K. & Lim, G. H. 2007 A study of hydraulic seal integrity. *Mech. Syst. Signal Process.* **21**, 1115–1126.

de Almeida, F. C. L., Brennan, M. J., Joseph, P. F., Dray, S., Whitfield, S. & Paschoalini, A. T. 2015 Measurement of wave attenuation in buried plastic water distribution pipes, Strojniški Vestn. *J. Mech. Eng.* **60**, 298–306.

Ferrante, M. 2011 Experimental investigation of the effects of pipe material on the leak law: leak in a steel pipe. *J. Hydraul. Eng.* **138**, 736–743.

Ferrante, M. 2012 Experimental investigation of the effects of pipe material on the leak law: leak in a steel pipe. *J. Hydraul. Eng.* **138**, 736–743.

Ferrante, M., Massari, C., Brunone, B. & Meniconi, S. 2011 Experimental evidence of hysteresis in the head-discharge relationship for a leak in a polyethylene pipe. *J. Hydraul. Eng.* **137**, 775–780.

Ferrante, M., Massari, C., Todini, E., Brunone, B. & Meniconi, S. 2013 Experimental investigation of leak hydraulics. *J. Hydroinform.* **15**, 666–675.

Fox, S. 2016 *Understanding the Dynamic Leakage Behaviour of Longitudinal Slits in Viscoelastic Pipes*. University of Sheffield, Sheffield.

Gao, Y., Brennan, M. J., Joseph, P. F., Muggleton, J. M. & Hunaidi, O. 2005 On the selection of acoustic/vibration sensors for leak detection in plastic water pipes. *J. Sound Vib.* **283**, 927–941.

GPSUK 2014 *GPS PE Pipe Systems UK*. Available from: www.gpsuk.com/articles/3/19/plastic-pipes%7Bnumber-one-choice-for-water%7C%0Asewerage-applications.html (accessed 19 December 2014).

Huang, N. E., Shen, Z., Long, S. R., Wu, M. C., Shih, H. H., Zheng, Q., Yen, N.-C., Tung, C. C. & Liu, H. H. 1998 The empirical mode decomposition and the {H}ilbert spectrum for nonlinear and non-stationary time series analysis. *Proc. R. Soc. Lond. A* **495**, 903–995.

Hunaidi, O. & Chu, W. T. 1999 Acoustical characteristics of leak signals in plastic water distribution pipes. *Appl. Acoust.* **58**, 235–254.

Kacur, J. & Jurecka, M. n.d. Speech Detection Using High Order Statistic (Skewness). Unpublished research.

Kaewwaewnoi, W., Prateepasen, A. & Kaewtrakulpong, P. 2007 A study on correlation of AE signals from different AE sensors in valve leakage rate detection. *ECTI Trans. Electr. Eng. Electron. Commun.* **5**, 113–117.

Mandic, D. P., Ur Rehman, N., Wu, Z. & Huang, N. E. 2013 Empirical mode decomposition-based time-frequency analysis of multivariate signals: the power of adaptive data analysis. *IEEE Signal Process. Mag.* **30**, 74–86.

Muggleton, J. M. & Brennan, M. J. 2004 Leak noise propagation and attenuation in submerged plastic water pipes. *J. Sound Vib.* **278**, 527–537.

Ofwat 2002 *Best Practice Principles in the Economic Level of Leakage Calculation*. Available from: www.ofwat.gov.uk/publications/commissioned/rpt com tripartitestudybstpractprinc.pdf (accessed 6 August 2014).

Pachoud, S., Gong, S. & Cavallaro, A. 2009 Space-time audio-visual speech recognition with multiple multi-class Probabilistic Support Vector Machines. In: *International Conference on Audio-Visual Speech Processing*, pp. 155–160.

Pal, M. 2008 *Leak Detection and Location in Polyethylene Pipes*. Doctoral thesis.

Papastefanou, A. 2011 *An Experimental Investigation of Leak Noise From Water Filled Plastic Pipes*. Doctoral thesis. Available from: http://eprints.soton.ac.uk/190853/.

Picone, J. 1993 Signal modeling techniques in speech recognition. *Proc. IEEE* **81**, 1215–1247.

Prime, M. & Shevitz, D. W. 1996 Linear and nonlinear methods for detecting cracks in beams. In: *Proc. 14th International Modal Analysis Conference*, Dearborn, Michigan, pp. 1437–1443.

Puust, R., Kapelan, Z., Savic, D. A. & Koppel, T. 2010 A review of methods for leakage management in pipe networks. *Urban Water J.* **7**, 25–45.

Sheng, J., Dong, S., Liu, Z. & Gao, H. 2016 Fault feature extraction method based on local mean decomposition Shannon entropy and improved kernel principal component analysis model. *Adv. Mech. Eng.* **8**, 1–8.

Su, Y., Jelinek, F. & Khudanpur, S. 2007 Large-scale random forest language models for speech recognition. *Lang. Speech* **1**, 598–601.

Sun, J., Xiao, Q., Wen, J. & Zhang, Y. 2016 Natural gas pipeline leak aperture identification and location based on local mean decomposition analysis. *Meas. J. Int. Meas. Confed.* **79**, 147–157.

Tayefi, P. 2014 *The Fatigue Response of Electrofusion Joints When Subject to Contamination*. Doctoral thesis, University of Sheffield.

UKWIR 2008 *National Sewers and Water Mains Failure Database (08/RG/05/26)*. Technical Report, UKWIR. London, UK.

Van Zyl, J. E. & Cassa, A. M. 2014 Modeling elastically deforming leaks in water distribution pipes. *J. Hydraul. Eng.* **140**, 182–189.

van Zyl, J. E., Alsaydalani, M. O., Clayton, C. R. I., Bird, T. & Dennis, A. 2013 Soil fluidisation outside leaks in water distribution pipes – preliminary observations. *Water Manag.* **166**, 546–555.

Varma, S. & Simon, R. 2006 Bias in error estimation when using cross-validation for model selection. *BMC Bioinform.* **7**, 1–8.

Wassaf, W. A., Bassim, M. N., Houssny-Emam, M. & Tangri, K. 1985 Acoustic emission spectra due to leaks from circular holes and rectangular slits. *J. Acoust. Soc. Am.* **77**, 916–923.

Water Services Association of Australia 2012 *Failure Modes in Pressurised Pipeline Systems*. Available from: www.sswm.info/sites/default/files/reference_attachments/WSAA%202003%20Common%20Failure%20Modes%20in%20Pressurised%20Pipeline%20Systems.pdf (accessed 23 December 2016).

Wu, Z. & Huang, N. E. 2005 Ensemble empirical mode decomposition: a noise-assisted data analysis method. *Adv. Adapt. Data Anal.* **1**, 1–41.

First received 25 September 2017; accepted in revised form 23 February 2018. Available online 2 April 2018