Al-Quayed *et al. J Wireless Com Network* (2020) 2020:256
https://doi.org/10.1186/s13638-020-01878-z

EURASIP Journal on Wireless Communications and Networking

RESEARCH Open Access

Lightweight feature extraction method for efficient acoustic-based animal recognition in wireless acoustic sensor networks

Fatima Al-Quayed, Adel Soudani* and Saad Al-Ahmadi

*Correspondence:
asoudani@ksu.edu.sa
Department of Computer Science, College of Computer and Information Science, King Saud University, P. O. Box 51178, Riyadh 11543, Saudi Arabia

Abstract

Wireless acoustic sensor networks represent an attractive solution that can be deployed for animal detection and recognition in a monitored area. A typical configuration for this application would be to transmit the whole acquired audio signal through multi-hop communication to a remote server for recognition. However, continuous data streaming can cause a severe decline in the energy of the sensors, which consequently reduces the network lifetime and questions the viability of the application. An efficient solution to reduce the sensor's radio activity would be to perform the recognition task at the source sensor then to communicate the result to the remote server. This approach is intended to save the energy of the acoustic source sensor and to unload the network from carrying, probably, useless data. However, the validity of this solution depends on the energy efficiency of performing on-sensor detection of a new acoustic event and accurate recognition. In this context, this paper proposes a new scheme for on-sensor energy-efficient acoustic animal recognition based on low-complexity methods for feature extraction using the Haar wavelet transform. This scheme achieves more than 86% in recognition accuracy while saving 71.59% of the sensor energy compared with the transmission of the raw signal.

Keywords: Wireless acoustic sensor networks, Acoustic-based recognition, Low-complexity feature extraction, Energy efficiency, In-network processing

1 Introduction

Many animal species face habitat loss in a rapidly changing world due to global warming, harmful human activities, etc. [1]. Hence, monitoring the endangered species in the ecosystems becomes an urgent worldwide concern. In this context, the automatic animal surveillance system presents an effective substitute for the ecologists' manual observations since they require costly and time-consuming on-site monitoring that can be infeasible depending on the monitored area's size.

The rapid progress in micro-electro-mechanical systems (MEMSs) has allowed the integration of low-cost micro-acoustic sensing components within network nodes, enabling the development of wireless acoustic sensor networks (WASNs). This technology enables a wide variety of unassisted acoustic-based monitoring applications for both

indoor and outdoor environments. Examples of these applications are endangered animal tracking [2], vehicle monitoring [3], speech localizing [4]. These applications rely, fundamentally, on recognizing the acoustic event of the target-of-interest that appears in the monitored area. In practice, the acoustic sensing deals with large volumes of data since it acquires audio signals under relatively high sampling frequencies [5]. Therefore, transmitting the whole raw audio signal via multi-hop communication for centralized recognition at the sink node is impractical [6]. As energy consumption is proportional to the volume of data transmitted by the node's wireless link, the mode of continuous data streaming can cause a fast energy decline of sensors involved in the radio communications, which shortens the network lifetime [6]. Besides, the time required to deliver all the relevant raw data to the remote recognition system can be significant, causing a long delay in recognizing the detected acoustic object [7]. Despite that the compressed sensing can be an effective solution for data reduction during signal acquisition [8]; it produces characteristics that are only useful for signal regeneration but not recognition using complex algorithms, which is infeasible at WASN due to the limited resources [9].

Alternatively, an energy-efficient solution would be to report only the result of the acoustic signal recognition (i.e., animal type) to the sink node with limited data size. Accordingly, the energy consumption of the wireless link, involved in radio communication, will be substantially reduced, contributing to extend the network lifetime [6]. More importantly, this approach avoids loading the network with unnecessary traffic, resulting in enhancing the network performance and availability. In this approach, recognizing the acoustic target that emits the acoustic signal is performed locally at the sensor node level. However, the validity of this solution depends on the efficiency of the designed WASN-based acoustic event recognition algorithms, where a balance between application's requirements and sensors' capabilities should be satisfied [6]. In this context, the effectiveness of the acoustic recognition systems is mainly based on the deployed feature extraction methods and the classification technique. Although there were several solutions proposed for acoustic-based target recognition for WASN [3, 4, 10–14], they suffer from high computational complexity due to the complicated feature extraction techniques (i.e., Fourier transforms) and the used classification approaches. Besides, these proposed solutions require considerable computational operations and need enough memory storage space to extract and store the large number of features.

Unlike Fourier transforms commonly used in recognition applications, discrete wavelet transforms (DWT), particularly the Haar wavelet transform, provide an efficient acoustic signal analysis with significantly reduced mathematical operations required for feature extraction [15]. Although wavelet transforms, in computer-based animal sound recognition, have been commonly studied [2, 10, 16–18], very few research work has addressed the application of DWT's features for acoustic animal recognition in WASN; see, for example, [10]. Despite the interesting results shown in this work [10], this approach requires applying complex algorithms for the classification and features optimization tasks, which cannot be executed at low-resources motes.

In our previous work [19], we studied the efficiency of on-sensor acoustic target recognition and localization using time-domain features. Despite the low-complexity of the proposed methods for feature extraction, the results have shown a considerable overlapping of the extracted features for different animal classes, which increases the difficulty

of classification. To solve this problem, we proposed a multi-label classification method in which the detected acoustic event is classified into the closest two classes of animals. The end-user notification carries, among other information, the extracted features for further tuned classification. Although the classification results were interesting, the proposed scheme was not capable of mapping the detected sound to a unique specific animal class.

This paper proposes a new scheme for energy-efficient acoustic animal recognition in WASNs with the goal of reducing the energy consumed during radio communication. To this end, the recognition task should be performed at the sensor node and thus, only the recognition result is delivered to the remote server instead of streaming the whole raw signal. The major challenges lie in selecting low-complexity accurate signal processing algorithms for feature extraction and classification tasks practical for the implementation on the sensor node. For this purpose, the Haar wavelet is used to derive a set of lightweight yet powerful features for on-sensor object recognition using a low-complexity classification method capable of differentiating between animal classes. The success criterion of this scheme is its capability to extend the lifetime of sensors' batteries while ensuring a successful animal classification. The rest of the paper is organized as follows. In the next section, we review the related work. Then, the proposed approach for acoustic sensing is described. After that, the paper details the experimentation and performance analysis in terms of system recognition accuracy and energy efficiency before concluding and highlighting future work.

2 Related work

An acoustic-based target recognition involves extracting features from the acquired signal to identify the source. Most of the existing recognition schemes have combined features from multiple domains; time, frequency, cepstrum, and wavelets to achieve a high recognition rate [20]. Cepstrum features that are successfully used for speech recognition were adopted for non-speech sounds in [21] and [22] to recognize bird and frog species' sounds, respectively. Frequency and time features can collectively achieve similar performance to cepstrum and sometimes even better, especially in terms of execution time [15]. The authors of [23] have examined the performance of a fusion of frequency, time, Mel, and Linear Frequency Cepstral Coefficients (MFCC's and LFCC's) on the recognition rate. This approach is tested on 199 classes of frog calls and achieved an accuracy of 95% using Support vector machines (SVMs). In [24], the proposed scheme was implemented to recognize nine frog species' sounds using a mixture of six spectral and temporal features. In [25], an optimal-performance approach used a wide range of features extracted from several domains to classify 22 classes of frog calls using five classification algorithms. Although the use of a combination of features increases the recognition accuracy of the approaches mentioned above, these approaches suffer from high time and space complexities. Hence, they are inappropriate for resource-constrained environments.

Due to the cost-intensive algorithms used in the domain of signal processing, only a little attention was given to acoustic-based target recognition in WASNs [3, 4, 10–14, 19]. In [14], two spectral features were used in the proposed approach for binary classification of the sound to either speech or music. The adopted feature

extraction algorithms are computationally complex, limiting their deployment for low-resources motes. Similarly, the approaches proposed in [3, 4, 11, 12] are characterized by their intensive-computations required for either spectral feature extraction or the used classifier. Thus, they are not adequate for tiny embedded devices. In [13], a featureless approach was designed for classifying two datasets: frog and cricket calls based on the sparse representation of signals to be suitable for WSN. Although the frog calls dataset's classification was successful, the classification performance of the cricket species dataset using the same technique was unsatisfactory. Due to the inconsistent results, this approach cannot be generalized to recognize sounds in a dataset of several classes of animals as intended by our proposed scheme.

Features derived from wavelet were adopted in several applications for the purpose of recognition in [2, 10, 16–18]. The proposed approach in [17] had achieved 78% and 96% using SVM and MLP, respectively, for bird sounds classification based on wavelet packet decomposition (WPD) transform. Similarly, a recent efficient WPD-based approach for bird sounds recognition and denoising was proposed in [18]. In [2], WPD-based features were used to classify frog sounds. However, WPD-based approaches require high computational resources to decompose the signal into deep levels for feature extraction. Therefore, Discrete wavelet transform (DWT) is preferred since it requires fewer operations than WPD. In [16], the raw wavelet coefficients and 26 features from cepstrum and spectral domains were merged to get high performance with a dataset of four animal classes. However, when DWT coefficients were used alone, poor results were obtained with a 25% recall due to the low-level meaning of wavelet coefficients that cannot be used directly as features. A well-known DWT-based approach used to recognize three datasets with a low number of classes per each dataset was [26]. The recognition performances ranged from 53 to 82% for the three datasets using 45 features extracted from 12 levels. We can conclude that these approaches require calculating a high number of features or coefficients. In [27], the ECG signals were classified using four decomposition levels to get a performance of 74%. Similarly, features derived from the Haar wavelet transform provided accurate results with low-complexity computations for the epilepsy episodes detection in the EEG signals [28]. In contrast to the binary classification of the signals in [27, 28], our proposed scheme is supposed to classify sounds as one of 12 classes.

To the best of our knowledge, DWT-based features proposed for acoustic recognition in WASN have been only addressed in [10]. Although the frog sounds accurately classified in [10], its efficiency is based on the application of computationally intensive algorithms for feature optimization and complex classifiers (i.e., DNN and genetic algorithm) that are not suitable to be deployed in low-resources motes. Moreover, the dataset used in this approach contains nine classes with only 49 syllables, which implies that the risk of over-fitting might occur. In summary, the current methods suffer from the need for high memory storage space and high computational complexity. Additionally, the energy consumption of the existing WASN-solutions has not been measured to evaluate the capability of these approaches in saving the sensor energy and thus extending network lifetime.

3 Methods

Wireless acoustic sensor network (WASN) consists of a group of acoustic sensors distributed over a specific area to exchange information wirelessly with the end-user at the sink node. This network type is mainly used for remote monitoring of acoustic targets appearing within the sensors field [2]. The WASN can be organized into several clusters by adopting an appropriate energy-efficient clustering algorithm, which is beyond the scope of this paper. Each cluster is composed of different member nodes and one cluster head (CH). The proposed solution provided in this paper aims to detect and recognize an acoustic object locally at sensors nodes whenever an object enters their sensing range. We assume that only one acoustic target appears at a specific time and all sensors are stationary with known positions.

To accomplish the target recognition task, first, the signatures of the targets-of-interest need to be stored in the sensor nodes' memory. During the network setup phase, the sink node broadcasts the signatures vectors of all classes to all CHs included in the network. Then, each CH will, in turn, send them to the associated member nodes. At the application runtime, every sensor node will acquire acoustic samples periodically to detect whether an acoustic event has occurred. This requests sensors to monitor any significant variation in the intensity of the sampled signal and whenever this variation exceeds a specific pre-determined threshold (RMS_0), an acoustic event is considered detected. Otherwise, the acoustic signal is discarded, and the acoustic sensor will resume its periodic sensing to detect new events. To decrease the notification packets' rate in the network, the sensor node, which detected the acoustic event, has to recognize the acoustic target's type before notifying CH. Consequently, sending unnecessary notifications to the CH in the case of false detections are avoided (i.e., false positives, where an acoustic object is detected, but it appears irrelevant). Unlike transmitting the whole raw acoustic signal, this approach decreases the packet size and the energy consumed during transmission. After detecting the presence of a new acoustic event in the sensors' vicinity, the on-board processor at each triggered sensor node will extract a set of features from the sampled acoustic signal. This set of features will define a vector (F) that represents the detected target. As the communication bandwidth is limited in WASN, a local pattern-based classification algorithm at each sensor will decide on the detected animal's type instead of transmitting the feature vector through the network. In other words, the extracted vector (F) is compared with all signatures, loaded in the memory of the sensors, using a suitable similarity measure to identify the object. The new acoustic target will be assigned to a specific class based on the highest similarity score between the signature of the corresponding class and the extracted vector (F). Once the object is recognized, the sensor node will send a notification message that contains the recognition result to the corresponding CH. Since all the triggered sensors will send their recognition decisions to the CH, a distance-rejection technique must be adopted at CH to consider only the most reliable result [29]. Accordingly, the result of the closest node to the detected target is assumed to be the reliable one [3]. This strategy avoids advertising false detections or redundant data to the sink node, which are unbeneficial for the end-user.

The fundamental steps required for the recognition of an acoustic sound in WASN are sampling, acoustic event detection, signal preprocessing, feature extraction, object

classification, and end-user notification. Note that the signal processing, feature extraction, and classification functions are performed in the microcontroller of every triggered sensor node. As previously discussed, not all feature extraction methods and classification algorithms can be performed at sensor nodes due to their limited processing resources. For adequate sensor-based implementation, the designed tasks should be capable of recognizing the target accurately at low-complexity computations, which allows per-node low-energy consumption. In the following sections, the required steps for acoustic-based recognition, depicted in Fig. 1, are detailed.

3.1 Sampling

Sampling all raw data samples of the acoustic signal at 8 kHz provides a satisfactory balance between the acoustic signal quality and the constrained resources [14]. Hence, using 8 kHz for sampling rate or lower is a common choice used in most of the WASN's acoustic recognition systems (see Table 4 for more details). Concerning the sampling period, careful selection is crucial since sampling can cause high energy consumption [30]. In our approach, the sampling period should exceed the minimum time required for the signal processing to conserve sensors' energy. This means that $T_s > T_\mathrm{p}$ must be satisfied, where T_s denotes the sampling period and T_p denotes the minimum time required to process the scheme including all the steps of signal acquiring, object detection, target recognition and CH notification. The sampling periods of 5, 10, and 20 s are common choices in animal recognition applications [31].

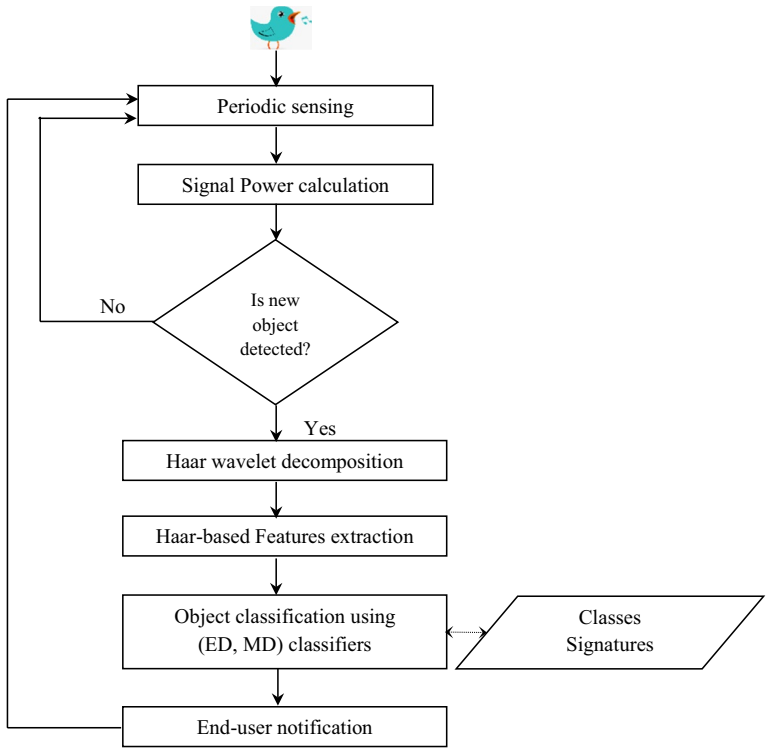

Fig. 1 Flowchart diagram of the proposed sensor-based acoustic animal recognition in WASN

3.2 Detection of a new acoustic signal

This step is performed to differentiate between real acoustic events from background noise regardless of whether this acoustic event is irrelevant or not, which will be decided at a later stage. In this step, the sensors in each cluster perform periodic sensing at a time interval (t) to acquire a new acoustic signal. These sensors should be aware of the acquired signal's power-level variation to decide whether a positive detection has occurred. For this purpose, the sensor node will calculate the root mean square (RMS) of the samples to obtain the average signal power $P_i(t)$. Then, the $P_i(t)$ is compared with a pre-defined threshold RMS_0. The threshold value RMS_0 is dependent on the noise level at the areas where the sensor nodes are placed, which is determined during network deployment as a part of the calibration procedures [4, 11]. As higher ambient noise levels may trigger unnecessary detections, updating RMS_0 value during network operation avoids increasing the detection complexity and thus energy expenditure [32]. The decision on object detection is determined based on the following detection function (D):

$$D = \begin{cases} 1 & P_i(t) > \mathrm{RMS}_0 \\ 0 & P_i(t) \leq \mathrm{RMS}_0 \end{cases} \tag{1}$$

An acoustic event is detected ($D = 1$) if the average signal power in a specific time interval exceeds RMS_0. Consequently, the acquired samples of the signal N will be passed to the next step for further processing. Otherwise, the current value of RMS_0 will be updated considering the calculated average power of the newly acquired signal according to the following equation:

$$\mathrm{RMS}_0 := \mathrm{RMS}_0 + (P_i(t) \times \alpha) \quad \text{where } 0 \leq \alpha \leq 1 \tag{2}$$

The threshold value RMS_0 plays an essential role in segmenting the signal into a time-series of raw acoustic data blocks, which results in event detection. A prior calibration step is required to allow the sensors to discover the background noise automatically within their local environment where the network is deployed [30]. Thus, this would help in determining the threshold value RMS_0 based on the sensors' awareness of the noise level when no acoustic event arises. Thus, in this stage, each sensor will compute the (RMS) of multiple acoustic signals over a time interval (t). Then, the average signal power will be used to determine the background noise threshold RMS_0. Self-calibration at sensors can be initiated automatically at any time during network operation for updating the current calibration procedures and to minimize the sensitivity to the variation of the noise level in the environment, if necessary, e.g., after orientation changes or an increase in the background noise.

3.3 Signal preprocessing

Signal preprocessing is a necessary step to prepare the signal for feature extraction. It involves framing and silence removal. In practice, acoustic signals are not stationary, but they can be considered stationary if they are analyzed for a short span of time. Therefore, in our approach, the acoustic signal is portioned into frames of 1024 samples that correspond to 0.128 s with an overlapping of 25% between consecutive frames to preserve the information contained in the boundaries of frames. Each frame is then multiplied

by a Hamming window before extracting features. The frame size used in our solution represents a reasonable choice since typical sensors (i.e., Mica/Mica2 series motes) have 4 Kb RAM only. As every sample point requires 2 bytes, this means a frame of 1024 data points allocates 2048 bytes of the sensor's RAM space while the remaining space is assigned to other modules [5].

Typically, acoustic animal records comprise periods of silence that need to be discarded when processing the signals. Thus, each frame is checked if it is silent to pass only the active frames to the feature extraction step. The primary purpose of this step is to emphasize the essential information contained in each sound record. Thus, only prominent non-silent frames of the sound signal are guaranteed to remain. Consequently, not only the quality of the feature extraction process is improved, but also the associated processing time is decreased. Since the background noise level might vary, the silence threshold for each sound record is calculated adaptively based on the signal's average energy. In our approach, we defined the silent frame as a frame whose Root-mean-square value (RMS_i) is below 10% of the long-term average signal energy [19]. Figure 2a–c shows the result of a sound record's silence removal algorithm in the dataset. The overlapping frames in Fig. 2b indicate the active parts that are remained for further processing. The silence removal method has the following steps:

1 Compute the threshold value T_{Silence} of each sound record based on the value of the average power of the whole signal $P_i(t)$ determined during the object detection phase (3):

$$T_{\mathrm{Silence}} = P_i(t) \times 0.10 \tag{3}$$

2 For each frame i, calculate the corresponding RMS_i value (4):

$$\mathrm{RMS}_i = \sqrt{\frac{\sum_{i=1}^{N} x_i^2}{N}} \tag{4}$$

where x_{i} is the value of the i^{th} sample and N represents frame length.

3 Compare the RMS_i value against T_{Silence}. If the RMS_i of frame i is below the signal threshold T_{Silence}, it is considered a silent frame and will be discarded, and the subsequent frame in line will be processed. Otherwise, frame i is considered an active frame and will be passed to the feature extraction step. This process is continuing until all frames of the signal are preprocessed. The function of the silent frames detection can be defined as expressed in (5):

$$S = \begin{cases} 1, & \mathrm{RMS}_i > T_{\mathrm{Silence}} \\ 0, & \mathrm{RMS}_i \leq T_{\mathrm{Silence}} \end{cases} \tag{5}$$

3.4 Feature extraction

The acoustic-based target recognition applications depend fundamentally on feature extraction methods to classify the detected acoustic event. Feature extraction methods should generate a compact representation of the features at a low computational cost to cope with WASN constraints while providing accurate classification results.

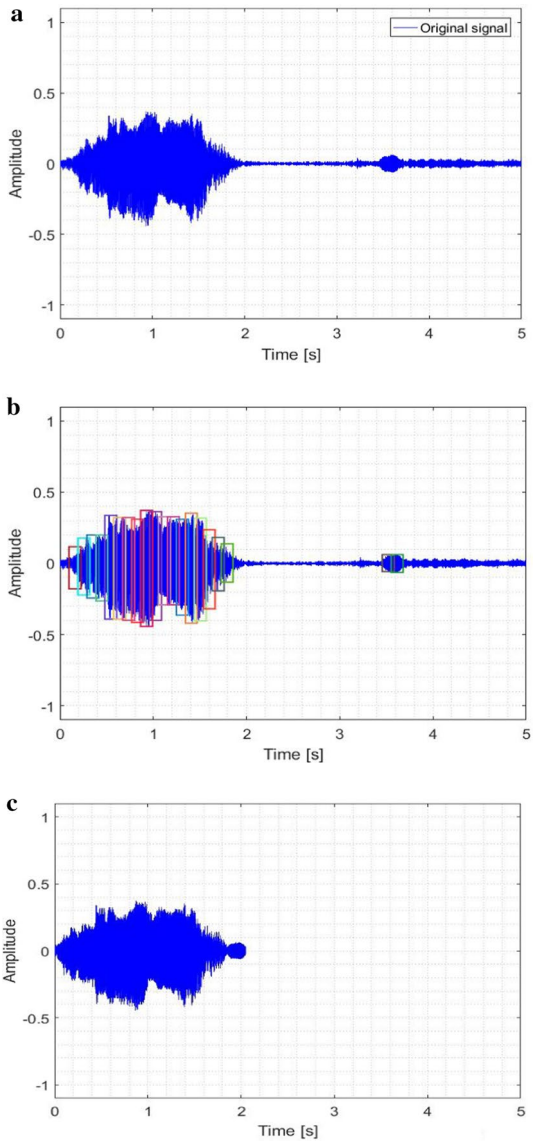

Fig. 2 The result of a 'cow' sound record preprocessing. **a** before silence removal, **b** the overlapping frames of the active part of the sound record, **c** the sound record after silence removal

Simple statistical measures extracted from the raw signal, such as mean, standard deviation, peak, variance, skewness, and kurtosis, were used for the acoustic target recognition, as stated in [33]. However, they failed to provide high accuracy in the results of our preliminary experiments and thereby, they are not part of the addressed features.

Although temporal domain features have a low time complexity of $O(N)$, they suffer from low discrimination ability between acoustic objects; thus, they are commonly combined with spectral features [23, 25, 34]. Even though this strategy increases the classification accuracy, it increases the computational complexity, which questions its practicality for the sensor-based implementation. More specifically, the extraction

of spectral features requires first transforming the signal into the frequency domain using Fast Fourier transform (FFT) or Short-Time Fourier transform (STFT), which have the complexity of $O(N \log_2 N)$ and additional $O(N)$ operations to compute the required feature. Moreover, STFT suffers from limited resolution in the time–frequency plane due to the fixed analysis window used for all frequencies [26]. Therefore, a better choice would be to analyze the range of different frequencies dynamically using a varied-size window to extract time–frequency features. This concept of multiresolution analysis is provided by the wavelet transform. Unlike Fourier transforms, wavelets can capture discontinuities and sharp spikes more efficiently due to wavelet functions' finite duration nature [35]. Capturing such information is significant in the differentiation between sounds. Compared to the wavelet packet decomposition (WPD) transform, discrete wavelet transform (DWT) is considered computationally inexpensive, particularly, the Haar wavelet that provides a robust signal analysis with low time complexity of $O(N)$ [36]. Therefore, we used the discrete wavelet transform (DWT) to decompose the signal into approximation (A) and detail (D) coefficients. In our preliminary tests, we evaluated the classification performances of different wavelets. The evaluated waveletes are the Haar, Db2, Db4, Db10, Sym2, Sym4, Sym8, Sym10, and Coif1, which are used in acoustic-based signal classification in [37]. We noticed that there is insignificant impact of the chosen mother wavelet on the classification rate which is consistent with findings in [27]. Therefore, we selected the computationally inexpensive and fast Haar wavelet transform that allows low-cost implementation at the sensors [36].

While raw wavelet coefficients can be employed for distinguishing between animal classes, our preliminary experiments have shown their inefficiency due to their low-level expressiveness and large dimensionality. Hence, for the selected decomposition level (l), the detail coefficients were extracted for each frame that composes the signal. After that, the mean of all frames' coefficients is calculated to get only one global vector of the detail coefficients per level. Then, we applied statistical measures on the detail coefficients to obtain compact features similar to what had been done in [26] by Tzanetakis. In [26], three features were used which are: the mean (***M***), standard deviation (***STD***) of the absolute values of the coefficients per subband, and the ratio of the mean values of two adjacent subbands (***RM***). In addition to these features, we extracted a set of wavelet-based features. Definitions and mathematical formulas of the extracted wavelet-based features are presented in "Appendix A". These features are the energy variance of coefficients in each subband (***E***) calculated in Eq. (A.1), temporal centroid per subband (***TC***) calculated in Eq. (A.2), the ratio of coefficients energy between two adjacent subbands (***ER***) calculated in Eq. (A.3), temporal centroids difference between two adjacent subbands (***TCD***) calculated in Eq. (A.4), and the Shannon entropy per subband (***P***) calculated in Eq. (A.5), which was proposed in [18]. We note that the ***ER*** and ***TCD*** features cannot be computed unless two decomposition levels are calculated. Finally, the values of each extracted feature are normalized to have zero-mean and unit-variance using $\overline{F} = \frac{F-\mu}{\sigma(F)}$ Where F and $\overline{F}$ are the original and normalized feature, respectively. μ and σ are the mean and standard deviation of the feature, respectively, obtained during the training phase.

The final decision on the features for implementation on WASN is based on two criteria: (1) their high accuracy to ensure efficient recognition and; (2) their adequacy for sensor-based implementation based on their computational complexity according to [15, 36]. Concerning the latter criterion, the time complexity of several implemented algorithms is presented in Table 1. Toward achieving the first criterion, we evaluated 14 features extracted from the wavelet, time, and frequency domains by adopting WEKA's attribute evaluator: *Gain Ratio*. The evaluated time and frequency-based features are energy (***g***), zero-crossing rate (***ZCR***), loudness (***L***), spectral roll-off (***SR***), spectral flux (***SX***), and spectral flatness (***SF***), which had been used in [23, 25, 34]. As a result, the descending list of the highest features in their gain ratio is: ***STD, E, P, L, g, ZCR, SX, SF, SR, TC, M***. In this experiment, we have applied one decomposition level for extracting wavelet features, thus, ***RM***, ***RE***, and ***TCD*** were not evaluated because they require two levels to be computed. The highest three features were wavelet-based features, and we found that [***Energy***, ***entropy***] have provided powerful discrimination between different animal classes contained in the dataset. More details about the recognition performance of different combinations are discussed in Sect. 4.1 while the energy evaluation of the selected wavelet features is validated through AVRORA simulator in Sect. 4.2.

3.5 Classification

The classification of the detected target is carried out in all triggered sensor nodes. However, CH adopts a distance-rejection technique to select the recognition result of the closest node to the detected target. In order to perform classification, the signatures of the animal classes are constructed offline and then broadcasted by the end-user to the deployed sensors during the configuration phase. These signatures are formed by computing the mean of the normalized features vectors for all the training samples per each class at the base station. Each signature is represented by $\text{signature}_C = \{\mu\text{feature}_1, \mu\text{feature}_2, \ldots\}$ where $C = 1, 2, .., 12$, which represents the class

Table 1 The computational complexity of feature extraction methods [15, 36]

Subset domain	Features (abbreviation)	Requires
Time	Energy(**g**)	$O(N)$
	Zero-crossing rate (**ZCR**)	
	Loudness (**L**)	
Frequency	Spectral flux (**SX**)	STFT which costs $O(N \log_2 N)$
	Spectral roll-off (**SR**)	
	Spectral flatness (**SF**)	
Wavelet	The mean of the subband's coefficients (**M**)	1-D Haar which costs $O(N)$
	The standard deviation of the subband's coefficients (**STD**)	
	The ratio of absolute mean values of two adjacent subbands (**RM**)	
	The energy variance of the subband's coefficients (**E**)	
	The energy ratio between two adjacent subbands (**ER**)	
	The temporal centroid per subband (**TC**)	
	The difference between the temporal centroids of two adjacent subbands (**TCD**)	
	The entropy per subband (**P**)	

number. In runtime, the features' vector is extracted from the newly detected acoustic signal. Then, the extracted vector is compared with all existing animal signatures using a lightweight distance-similarity classifier. More specifically, each sensor involved in the recognition computes the distances between the extracted feature vector (F) and every animal signature. Then, all these distances are stored in a distance vector $D_C = \{d_1, d_2, ..., d_{12}\}$, where d_1 represents the distance between the (F) and *class* 1 signature and so on. Then, the class label (C) of the unknown object will be determined based on the shortest value in D_C vector. Euclidean distance (ED) or Manhattan distance (MD) is two of the common similarity measures in the context of data mining and classification applications [20]. Definitions and mathematical formulas of Euclidean and Manhattan distance are presented in "Appendix B"; see B.1 and B.2.

Standard metrics used in audio-based recognition applications are recall, precision. In the following equations, TP, TN, FP, and FN stand for the correctly predicted positive instances, the correctly predicted negative instances, the actual negative instances that are classified incorrectly as positive, and the actual positive instances that are classified incorrectly as negative, respectively.

- Recall: this metric measures the true positive rate, which is critical for the system performance. The goal is to obtain a high recall score, which indicates that the system is not missing an animal of interest. This metric's high value means that we get a low miss rate (i.e., low number of undetected targeted animals), which is required for the system reliability.

$$\text{Recall} = \frac{TP}{TP + FN} \tag{6}$$

- Precision: it measures the positive predictive value that is significant for situations when the cost of the false positive detections (i.e., the detection of untargeted animals) is high. In our case, these instances represent the number of non-targeted animals that are recognized by our system. Thus, recognizing irrelevant sounds will trigger unrequired processing in the sensors, which should be avoided. The goal is to obtain a high precision value as it represents the classification predictions' accuracy per class.

$$\text{Precision} = \frac{TP}{TP + FP} \tag{7}$$

4 Results and discussion

In this section, we describe the set of experiments conducted to assess the proposed scheme performances in terms of recognition accuracy and energy efficiency. The performance analysis addressed the application level to measure the efficiency of the scheme in terms of classification accuracy of the detected sound based on the extracted wavelet features. For this purpose, we implemented the proposed scheme with MATLAB. We used a dataset containing audio records belonging to twelve different animals, namely: {Dog, Rooster, Pig, Cow, Frog, Cat, Chicken, Insect, Sheep, Crow, Cricket, and bird}. This set of animals' sounds represents a subset from a larger dataset used for environmental

sound classification known as ESC-50 [38]. The dataset records, which contain 480 files, are stored in.wav format with a sampling frequency of 44.1 kHz. Moreover, we collected for the evaluation of the proposed scheme 127 records at.wav format from different websites of animal sound libraries such as [39]. The collected records have different sampling frequencies with variable durations, and they often contain background noise. Therefore, the preprocessing of these records is an essential step to mitigate the issues mentioned earlier. 20 records out of 480 included in the ESC-50 dataset were excluded from the dataset due to the high noise level. We also performed frequency downsampling for the records to be at a unified sampling frequency of 8 kHz. The total number of the collected animal sound records is 587 records with durations ranging from 1 to 5 s. From these records, we have used 70% for training and 30% for testing.

4.1 The recognition accuracy of the proposed scheme

We evaluated the discrimination capability of wavelet-based features to determine their impact on animal recognition accuracy. First, we examined the recognition's performance of the extracted wavelet-based features, as shown in Table 2. Those features are derived from five decomposition levels of the acquired signal. Interestingly, features combinations derived from this decomposition level recorded the highest recognition accuracy compared to lower levels. However, the increment in the decomposition levels raises the computations and increases the features vector's size. Due to the limited memory on tiny devices, this number of features is voluminous and impractical for the application. Alternatively, we studied other combinations of the eight wavelet features to select the appropriate non-lengthy efficient feature vector that ensures a trade-off between complexity and recognition accuracy. For this purpose, the WEKA tool was used where gain ratio attribute evaluator and *Ranker* search methods were adopted to check the discrimination capability of different features' combinations. The recall results of different combinations of features, which showed the highest gain ratio, according to

Table 2 The recall results of different combinations of wavelet-based features

Set	Wavelet features selected	Average recall using ED (%)	Average recall using MD (%)	Number of features	Wavelet decomposition level(s)
#1	All features	88.82	89.86	37	5
#2	Entropy of level 2 Energy of level 1 Energy Ratio of level 2 & 1 Centroid of level 1	85.70	87.09	4	2
#3	Entropy of level 2 Energy of level 1 Energy Ratio of level 2 & 1	85.38	84.86	3	2
#4	STD of level 1 Energy of level 1 Entropy of level 1	86.04	86.87	3	1
#5	Energy of level 1 Entropy of level 1	86.08	85.59	2	1
#6	STD of level 1 Energy of level 1	85.03	85.44	2	1
#7	STD of level 1 Entropy of level 1	85.52	85.21	2	1

WEKA, are shown in Table 2. We noticed that higher decomposition levels or more features do not necessarily increase the recall performance. (e.g., set#3 are extracted from the second decomposition level and have more features than set#5 but recall results obtained with set#3 are lower than of those obtained with set#5).

We also found that decomposition levels one and two can produce high results using both Euclidean and Manhattan Distances. The number of features and the decomposition levels used for extracting each set are depicted in Table 2. Compared with the results obtained with the features set extracted from two levels reported in Table 2, we noticed that a satisfactory recognition rate could be obtained with only one decomposition level and with a smaller set of features. The reason is that the detail coefficients of the first level capture the high frequencies and abrupt changes of the signal that help in differentiating between animal classes, which are essential for the recognition task.

Using set #2 and set #4 of features achieved, the highest recall results compared with the rest, particularly with set #5 of features. However, calculating set #2 and set #4 of features needs higher computations and higher processing resources to decompose the signal at a higher level and to extract more features. Indeed, despite the higher complexity of the methods used to extract set #2 and set #4 of features, the gain in recognition accuracy was improved by only around 1%, which is not worthwhile. Interestingly, the method used to extract set #5 of features requires fewer computational operations to decompose the signal into only one decomposition level and to calculate two features while guaranteeing high recognition accuracy. Thus, we considered feature set #5 an adequate choice for the proposed scheme since it guarantees a satisfactory compromise between sound recognition capability and energy expenditure.

We evaluated the proposed recognition scheme's recall and precision based on set#5 of features using Manhattan Distance and Euclidean Distance classification methods for the set of sound record of animals. The results are shown in Fig. 3.

In terms of precision metric, we noticed that some classes' classification results based on ED are slightly better than using MD. Figure 3 shows that 100% of audio records for seven classes were correctly classified according to the recall results. These classes are Rooster, Frog, Cat, Chicken, Insect, Crow, and Crickets. Based on human listener classification, the recall rates of individual classes such as rooster, frog, cat, chicken, crow, and crickets were below 80% according to [38]. This indicates that our proposed scheme can accurately classify the acoustic objects that belong to these classes when they appear

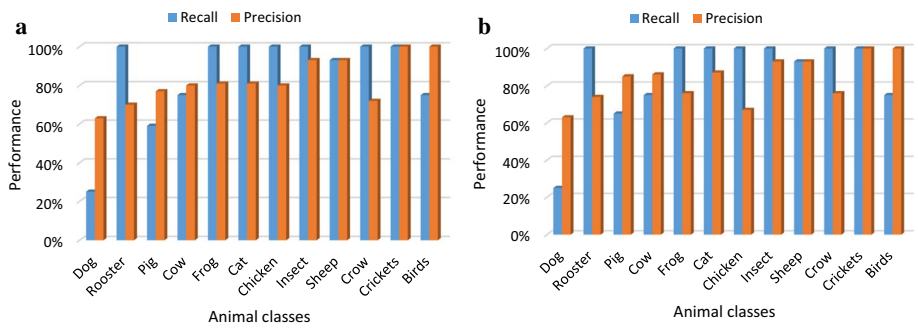

Fig. 3 The recall and precision results of set #5 of features using **a** Manhattan distance, **b** Euclidean distance

in the area of surveillance. We can also note that the proposed scheme produced 80% or more in terms of precision for seven classes using Euclidean distance. This high precision performed by the proposed scheme would reduce false detection, making it suitable for an extended range of applications. The proposed approach achieved an average recall of 86.08% using Euclidean Distance, where seven animal classes were 100% recognized, which demonstrates the high accuracy of our approach. However, the recall results of *Dog* and *Pig* classes using this scheme are low due to the wide variety of waveforms contained in the records of these classes. In fact, the *dog* class contains records of puppies' sounds while *Pig* class includes records of snort, grunt, squeal, and oinks sounds. Hence, the classification model could not accurately learn the mapping between classes and different waveforms within one class, especially with the limited number of records per class. We believe that the performance of the proposed approach of animal recognition can be further improved if a larger dataset of animal sounds is used in the learning phase.

The mean distributions of the selected features: energy and entropy of the sounds of all animal classes are presented in Fig. 4a, b. These distributions show how these two features have distinguishable values per class. When combined, these features provide high accuracy for the recognition task.

Although set #5 of features are designed to be effective under sampling frequency of 8 kHz, they also showed an encouraging performance of recall using 16 kHz and 44.1 kHz, where 84.61% and 85.10% were achieved using Euclidean Distance, respectively. This result attests the scalability of our proposed scheme under different sampling frequencies.

4.2 The energy efficiency of the proposed recognition scheme implemented on the sensor

We studied the energy efficiency of the proposed scheme when implemented in a wireless acoustic sensor. For this purpose, we used the AVRORA tool [40], which is an instruction-level emulator of sensors platforms that allows the evaluation of energy consumed in executing internal algorithms and communicating data to a remote device. We estimated the energy consumption for MICAz motes based on ATmega128L microcontroller and using a RAM of 4 kB.

The energy consumption and execution time of the proposed scheme were evaluated based on three different scenarios, as depicted in Table 3. In the first scenario,

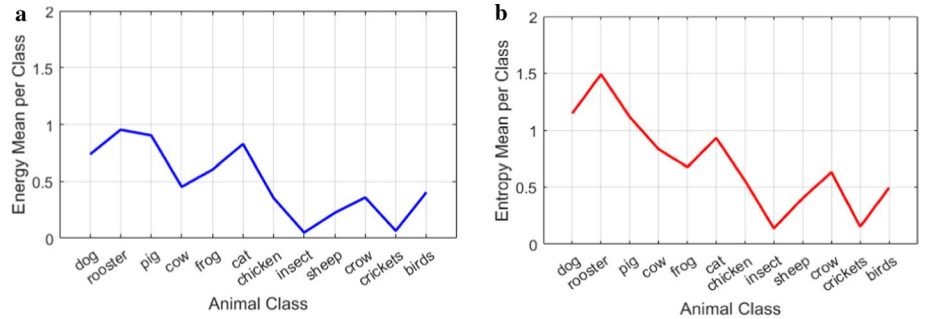

Fig. 4 The mean distribution of features **a** energy, **b** entropy

Table 3 Evaluation of the three processing scenarios at the sensor using MICAz

Measured attribute	Energy consumption (mJ)	Time (ms)
Scenario #1: The full scheme	37.05	500
Processing	36.79	496
Notification	0.26	4.50 E−3
Scenario #2: The scheme without classification	37.02	498
Processing	36.75	493
Notification (vector of features)	0.27	4.76 E−3
Scenario #3: Transmission of 1 s of the raw acoustic signal	130.42	1511.89

the sensor processes the whole tasks of the animal recognition scheme. Then, it transmits a notification packet of one byte that carries the detected class's identity to the end-user. In the second scenario, the acoustic sensor processes all the proposed steps except the classification, which will be performed at the server level. This means that the server will be notified with a packet containing the 2-D vector of floating-point features, where each feature is represented with 4 bytes. The server performs the classification task based on those received features. In the third scenario, the acoustic sensor will send the raw data samples to the remote server without any processing. Each sample of the signal is represented by 2 bytes. In our approach, the acoustic signals are assumed to be in a mono-format with a sampling frequency of 8 kHz (16-bit). The simulation results for processing an acoustic signal with a length of 1 s are illustrated in Table 1. As we can see from this table, the overall cost of processing the whole scheme (scenario #1) is around 37 mJ and needs around 0.5 s for execution. Compared to (scenario #2) where the classification is processed at the server-side, the gain in energy and time is negligible. This is because the classification method, proposed in our scheme, has a low-complexity and does not require intense processing. We think that implementing the classification at the sensor side would avoid the transmission of useless notification and would reduce the communication overhead [32]. When compared to (scenario #3), which corresponds to transmitting a record of 1 s of raw data of the acoustic signal, we can note that (scenario #1) saves 71.5% of sensor energy, proving the efficiency of our approach in extending sensor lifetime.

Table 4 sums a comparison with the performances and the characteristics of the most relevant similar solution designed for WASNs. The obtained results showed that our recognition scheme consumes more energy than the solution proposed in [20]. This is because extracting wavelet-based features requires a higher processing bandwidth than extracting features in the time domain. However, it is important to note that the scheme proposed in [20] was unable to perform single-label classification. In contrast, our scheme proved the ability to perform one-label animal recognition with a recall of 86% for twelve classes. In terms of execution time, it was reported in [5] that implementing a 512-point Fast Fourier Transform algorithm needs 15 s in ExScal motes. However, the proposed recognition scheme's execution needs only 0.5 s in the

Table 4 WASN-based animal recognition systems

	Approach	Features	Number of features	Classifier	Dataset (no. of classes) = no of files	Recognition accuracy (Recall)	Sampling frequency
1	Croker et al. [11]	Frequency and time	5	ED	Frog (5) = 100	Recall = 85% Accuracy = 89%	16 kHz
2	Dang et al. [12]	Envelope Extraction	Not specified	Matched filtering	Frog (3) = not specified	Accuracy = 90%	< 10 kHz
3	Wei et al. [13]	From Gradient Projection for Sparse Reconstruction	featureless using a sparse representation	Their own ι_1-minimization Sparse Approximation-based classifier	Frog (14) = 228 crickets (20) = 663	Recall ≈ 98% Recall ≈ 50%	24 kHz
4	Colonna et al. [10]	Wavelet	4	k-NN	Anurans(9) = 49 syllables	96.25% 94.16 86.96%	44.1 kHz 11 kHz 5.5 kHz
5	Algobail et al. [19]	Time	2	ED	Animals (7) = 114	81.34%	44.1 kHz
6	Our scheme	Wavelet	2	MD ED	Animals (12) = 587	85.59% 86.06%	8 kHz

MICAz mote to send the result to the remote server, making it suitable for real-time implementation in WASN.

Also, Table 4 shows that the approaches proposed in [10–12] were capable of providing higher recognition accuracy than our scheme. However, the energy efficiency of those schemes was not proved for the execution in limited-resources systems. In fact, feature extraction methods and complex classifiers used in these approaches are expensive in terms of computations and thereby impractical for WASN. Although the frog sounds dataset classification was accurate in [13], the cricket sounds dataset classification was unsatisfactory. This indicates that the proposed scheme of [13] is not scalable in classification performance for a wide range of animal classes. In contrast, our scheme accurately assured the recognition of one animal sound from a group of twelve animal classes.

5 Conclusions

This paper presented a lightweight scheme for acoustic-based animal recognition designed for sharply limited-resources systems (i.e., MICAz motes) seeking to extend their lifetime. Instead of streaming the whole raw signal of the acoustic event to the remote server, the proposed approach is intended to locally recognize the target and communicate only the recognition result to the server. Accordingly, the node energy will be saved, which contributes to extend the application's viability. The scheme's effectiveness is mainly based on applying low-complexity accurate signal processing algorithms for features extraction and classification adequate to be deployed on sensor nodes. For this purpose, we adopted low-complexity Haar-based features capable of ensuring a high granularity and accurate classification of the detected animal sound. The experimental results have shown that the proposed scheme ensures 86% of recognition recall saving 71.5% of energy compared to streaming the whole

acquired acoustic signal to the remote server. The results showed further that the whole Haar-based scheme's execution time is lower by orders of magnitude compared with FFT algorithms commonly used in similar approaches. We can conclude that this approach avoids loading the network with probably unnecessary traffic, leading to increased network performance and availability. Nevertheless, the development of an efficient low-energy monitoring system requires further research to be conducted regarding the design of low-complexity methods for object localization and tracking to build upon the current recognition system.

Abbreviations

WASN: Wireless acoustic sensor network; WPD: Wavelet packet decomposition; DWT: Discrete wavelet transform; FFT: Fast Fourier transform; EEG: Electroencephalogram; ECG: Electrocardiogram; DNN: Deep neural network; STFT: Short-time Fourier transform; Db: Daubechies; Sym: Symlet; Coif: Coiflet.

Acknowledgements

The authors extend their appreciation to the Deanship of Scientific Research at King Saud University for funding this work through research group No (RG-1439-023).

Authors' contributions

All authors were involved in the discussion of the work described in this paper. All authors contributed in writing the manuscript. All authors read and approved the final manuscript.

Funding

The authors extend their appreciation to the Deanship of Scientific Research at King Saud University for funding this work through research group No (RG-1439-023).

Availability of data and materials

The dataset used and analyzed during the current study is available in the ESC: Dataset for Environmental Sound Classification repository, https://github.com/karolpiczak/ESC-50 [38]

Competing interests

The authors declare that they have no competing interest.

Appendices

A. The extracted haar-based features

1 The energy variance of coefficients in each subband (***E***). This feature represents the frequency distribution, which is calculated using the following equation.

$$E = \frac{\sum_{i=1}^{N}(x(i) - \overline{X})^2}{N-1} \tag{A.1}$$

where $x(i)$ indicates the value of coefficient i, N represents the number of coefficients in level L, and $\overline{X}$ represents the mean of the coefficients' values.

2 The temporal centroid per subband (***TC***). It represents the temporal distribution of the signal, which is calculated as follows:

$$\mathrm{TC} = \frac{\sum_{i=1}^{N}\left(|x(i)|^2 t[i]\right)}{\sum_{i=1}^{N}|x(i)|^2} \tag{A.2}$$

where $x(i)$ indicates the value of coefficient i, N represents the number of coefficients in level L, t[i] represents the time index.

3 The ratio of coefficients energy between two adjacent subbands (***ER***). It represents the amount of change in a frequency distribution, which is calculated as follows:

$$E\,\text{Ratio} = \frac{E_L}{E_{L+1}} \tag{A.3}$$

where L corresponds to the decomposition level number, and E is computed according to Eq. (6).

4 The temporal centroids difference between two adjacent subbands (***TCD***). It represents the amount of change in temporal distribution, which is calculated as follows:

$$\text{TCD} = \text{TC}_{L+1} - \text{TC}_L \tag{A.4}$$

where L corresponds to the decomposition level number.

5 The Shannon entropy per subband (***P***). It represents the temporal distribution of the signal, which is calculated as $\boldsymbol{P} = -\sum_i p_i \log p_i$ where p_i is the probability of coefficient i appearing in the subband. However, we used a slightly different version of this equation used in [18] as follows:

$$\mathbf{P} = -\sum_i x(i)^2 \log x(i)^2 \tag{A.5}$$

Where $x(i)$ indicates the value of coefficient i. Shannon entropy was selected among many others (i.e., log-energy, threshold, and norm) due to its performance during preliminary experiments.

B. The mathematical formulas of Euclidean and Manhattan distance

Euclidean distance can be defined as a straight line between two points in Euclidean space. Suppose we have two features' vectors $f = (x_1, x_2, \ldots, x_n)$ and $f_C^{'} = (y_1, y_2, \ldots, y_n)$. in $\boldsymbol{n}$ -dimensional space of features. The Euclidean distance of the two vectors is the sum of the square difference for each feature. It is calculated using this formula:

$$d_{\text{E}}(f, f_{\text{C}}') = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \tag{B.1}$$

The Manhattan distance of the two vectors is the sum of the absolute differences of their features. It is calculated using this formula:

$$d_{\text{M}}(f, f_{\text{C}}') = \sum_{i=1}^{n} |x_i - y_i| \tag{B.2}$$

where f is the detected target vector, $f_C^{'}$ is the signatures of classes, and n represents the number of features.

Received: 21 February 2020 Accepted: 2 December 2020
Published online: 14 December 2020

References

1. J.G. Colonna, M. Cristo, M. Salvatierra, E.F. Nakamura, An incremental technique for real-time bioacoustic signal segmentation. Expert Syst. Appl. **42**(21), 7367–7374 (2015)

2. J. Xie, M. Towsey, J. Zhang, P. Roe, Adaptive frequency scaled wavelet packet decomposition for frog call classification. Ecol. Inform. **32**, 134–144 (2016)
3. M.F. Duarte, Y.H. Hu, Vehicle classification in distributed sensor networks. J. Parallel Distrib. Comput. **64**(7), 826–838 (2004)
4. Y. Guo, M. Hazas, Localising speech, footsteps and other sounds using resource-constrained devices, in *International Conference on Information Processing in Sensor Networks (IPSN)* (2011)
5. L. Gu et al., Lightweight detection and classification for wireless sensor networks in realistic environments, in *Proceedings of the 3rd International Conference on Embedded networked sensor systems* (2005)
6. G. Wittenburg, N. Dziengel, S. Adler, Z. Kasmi, M. Ziegert, J. Schiller, Cooperative event detection in wireless sensor networks. IEEE Commun. Mag. **50**(12), 124–131 (2012)
7. G. Wittenburg, N. Dziengel, A system for distributed event detection in wireless sensor networks, in *Proceedings of the 9th ACM/IEEE International Conference on Information Processing in Sensor Networks* (2010), pp. 94–104
8. D.L. Donoho, Compressed sensing. IEEE Trans. Inf. Theory **52**(4), 1289–1306 (2006)
9. P. William, Low complexity feature extraction for classification of harmonic signals, PhD thesis, University of Nebraska, 2011
10. J.G. Colonna, A.D. Ribas, E.M. Santos, E.F. Nakamura, Feature subset selection for automatically classifying anuran calls using sensor networks, in *The 2012 International Joint Conference on Neural Networks (IJCNN)* (2012), pp. 10–15
11. B. Croker, N. Kottege, Using feature vectors to detect frog calls in wireless sensor networks. J. Acoust. Soc. Am. **131**(5), 400–405 (2012)
12. T. Dang, N. Bulusu, W. Hu, Lightweight acoustic classification for cane-toad monitoring, in *Conference on Signals, Systems and Computers* (2008), pp. 1601–1605
13. B. Wei, M. Yang, Y. Shen, R. Rana, C. T. Chou, W. Hu, Real-time classification via sparse representation in acoustic sensor networks categories and subject descriptors, in *Proceedings of the 11th ACM Conference on Embedded Networked Sensor Systems* (2013)
14. B. Chen, Audio recognition with distributed wireless sensor networks, Master thesis, University of Victoria, 2010
15. E.L. Salomons, P.J.M. Havinga, A survey on the feasibility of sound classification on wireless sensor nodes. Sensors **15**(4), 7462–7498 (2015)
16. D. Mitrovic, M. Zeppelzauer, C. Breiteneder, Discrimination and retrieval of animal sounds. Master Thesis, Vienna University of Technology, 2005
17. A. Selin, J. Turunen, J.T. Tanttu, Wavelets in recognition of bird sounds. EURASIP J. Adv. Signal Process. **2007**, 1–9 (2007)
18. N. Priyadarshani, Wavelet-based birdsong recognition for conservation, PhD thesis, Massey University, 2016
19. A. Algobail, A. Soudani, S. Alahmadi, Energy-aware scheme for target recognition and localization in wireless acoustic sensor networks. Int. J. Distrib. Sens. Networks **15**(11), 1550147719891406 (2019)
20. N.C. Han, S.V. Muniandy, J. Dayou, Acoustic classification of Australian anurans based on hybrid spectral-entropy approach. Appl. Acoust. **72**(9), 639–645 (2011)
21. J. Cai, D. Ee, B. Pham, P. Roe, J. Zhang, Sensor network for the monitoring of ecosystem: bird species recognition, in *3rd International Conference on Intelligent Sensors, Sensor Networks and Information* (2007), pp. 293–298
22. C. L. Ting Yuan, D. Athiar Ramli, Frog sound identification system for frog species recognition, in *Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering*, vol. 109 (2013), pp. 41–50
23. J.J. Noda, C.M. Travieso, D. Sánchez-Rodríguez, Methodology for automatic bioacoustic classification of anurans based on feature fusion. Expert Syst. Appl. **50**, 100–106 (2016)
24. C.J. Huang et al., Intelligent feature extraction and classification of anuran vocalizations. Appl. Soft Comput. J. **19**, 1–7 (2014)
25. J. Xie, M. Towsey, J. Zhang, P. Roe, Acoustic classification of Australian frogs based on enhanced features and machine learning algorithms. Appl. Acoust. **113**, 193–201 (2016)
26. G. Tzanetakis, G. Essl, P. Cook, Audio analysis using the discrete wavelet transform, in *Conference in Acoustics and Music Theory Applications* (2001)
27. P. De Chazal et al., Using Wavelet Coefficients For The Classification Of The Electrocardiogram, in *Annual International Conference on IEEE Engineering in Medicine and Biology*, vol. 1 (2000), pp. 64–67
28. S. Alhassan, M. A. Aldammas, A. Soudani, Energy-efficient sensor-based EEG features' extraction for epilepsy detection, in *Current and Future Trends of Information and Communication Technologies in Healthcare (ICTH 2019)*, vol. 160 (2019), pp. 273–280.
29. N. Dziengel, M. Seiffert, M. Ziegert, S. Adler, S. Pfeiffer, J. Schiller, Deployment and evaluation of a fully applicable distributed event detection system in Wireless Sensor Networks. Ad Hoc Netw. **37**, 160–182 (2016)
30. N. Dziengel, Distributively observed events in wireless sensor networks, PhD thesis, Freien University Berlin, 2015
31. A.J. Garcia-Sanchez et al., Wireless sensor network deployment for monitoring wildlife passages. Sensors **10**(8), 7236–7262 (2010)
32. I.F. Akyildiz, W. Su, Y. Sankarasubramaniam, E. Cayirci, Wireless sensor networks: a survey. Comput. Netw. **38**, 393–422 (2002)
33. H. Kim, N. Moreau, T. Sikora, *MPEG-7 Audio and Retrieval: Audio Content Indexing and Retrieval* (Wiley, West Sussex, England, 2005)
34. S. Chu, S. Narayanan, C.-C.J. Kuo, Environmental sound recognition with time–frequency audio features. IEEE Trans. Audio. Speech. Lang. Process. **17**(6), 1142–1158 (2009)
35. K. Dziedziech, W.J. Staszewski, B. Basu, T. Uhl, Wavelet-based detection of abrupt changes in natural frequencies of time-variant systems. Mech. Syst. Signal Process. **64–65**, 347–359 (2015)
36. F. Germain, *The Wavelet Transform Applications in Music Information Retrieval* (McGill University, Montreal, 2009), pp. 1–29

37. K. Umapathy, S. Krishnan, R.K. Rao, Audio signal feature extraction and classification using local discriminant bases. IEEE Trans. Audio Speech Lang. Process. **15**(4), 1236–1246 (2007)
38. K.J. Piczak, ESC: Dataset for environmental sound classification, in *Proceedings of the 23rd ACM International Conference on Multimedia* (2015), pp. 1015–1018
39. Free Sound, FREE sound effects (2009), https://www.freesoundeffects.com/. Accessed 20 Feb 2020
40. B.L. Titzer, D.K. Lee, J. Palsberg, Avrora: scalable sensor network simulation with precise timing, in *2005 4th International Symposium on Information Processing in Sensor Networks, IPSN 2005* (2005)

Publisher's Note