# An energy-efficient WMSN-based system for endangered birds monitoring

Aya Sakhri, Moufida Maimour, Eric Rondeau, Noureddine Doghmane, Saliha Harize

# An Energy-Efficient WMSN-based System for Endangered Birds Monitoring ⋆

**Aya Sakhri** [*,**] **Moufida Maimour** [*] **Eric Rondeau** [*]
**Noureddine Doghmane** [**] **Saliha Harize** [**]

[*] *Lorraine University, CNRS, UMR 7039, Campus Sciences, BP 270239, F-54000 Nancy, France (e-mail:aya.sakhri@univ-lorraine.fr).*
[**] *LASA laboratory, Badji-Mokhtar Annaba University, Algeria*

**Abstract:** Recently, more attention is being payed to the ecological environment because of the rapid decline of animal species especially birds. Wireless Multimedia Sensor Networks (WMSN) can be leveraged to monitor, protect and assess changes in birds populations by capturing and sending images in real time. To deal with their limited bandwidth and energy resources, we propose to reduce the amount of visual data to report by sending only images of interest to the collect station. This can be achieved by identifying the endangered target species locally based on their calls before triggering the camera. However, one prominent obstacle is that in wilderness environment, it is difficult to obtain good quality audio signals. This is due to interfering environmental noises that can mask vocalizations of interest so the audio recognition is likely to fail. Therefore, we set up an automated birdsong recognition along with an appropriate noise reduction-based method in order to improve the audio classification accuracy. This ultimately leads to a better use of the network energy and avoids its waste in transmitting useless images.

*Keywords:* Birds Monitoring, WMSN, Audio Denoising, Audio Classification, Energy Efficiency.

## 1. INTRODUCTION

Birds are good bio-indicators of ecosystem biodiversity since their decline can cause serious problems to the environmental health and safety. The State of the World's Birds Allinson (2018) revealed that we lost more than 161 species, which represents an extinction rate much higher than natural. This is because of several important reasons including human activities and climate changes. Hence, the survival of our environmental system from various dangers is a primary responsibility which can be achieved through endangered birds monitoring in their natural habitats. Most of the time, the monitoring is done manually in a limited space and time. Moreover, the physical presence of observers in the field may disturb the birds behavior. This is why, we propose to use wireless multimedia sensor network (WMSN) Almalkawi et al. (2010) to build an automatic system for bird surveillance in their natural habitat. The ultimate goal is to identify and estimate the population of endangered birds species based on their transmitted photographs.

Nevertheless, WMSN are characterized by their limited resources in terms of energy, memory and bandwidth. Besides, the sensor network contains hundreds of nodes deployed in hostile environment, which makes the replacement of batteries a hard task. Therefore, improving the lifetime of the sensor nodes to ensure efficient network utilization is one crucial challenge to deal with. In fact, image processing and transmission are the major energy consuming stages in a WMSN-based surveillance system. It is therefore essential to reduce the amount of the visual data to capture and transmit. To address this issue, multiple low-cost visual data compression algorithms have been proposed ZainEldin et al. (2015). Also, Civelek and Yazici (2016) considered the extraction of relevant image features, others limited the image capture using motion detection as Magno et al. (2013). However, all these techniques are not sufficient since camera nodes continue to capture and process images including irrelevant ones which is just a waste of energy and resources.

In this paper, we propose to capture only relevant images of target species. This can be achieved by designing a low-complexity system based on local audio detection and recognition of target species in real time prior to triggering image capture. This is motivated by the fact that visual data processing is more expensive when compared to audio data that is of significantly lower size. This local audio identification allows to avoid the continuous processing and transmission of unnecessary images that can paralyze the entire sensor network. In addition, it provides further consolidation of the recognition accuracy. There are very few works that combine audio-visual data processing in a sensor node to decrease the energy consumption as proposed by Koyuncu et al. (2018). However, their model does not take into account the overlapping of various environmental noises, which is a major constraint in our application. It degrades the quality of the target sound Xie et al. (2021) which results in false identification of species.

⋆ This work was supported in part by the PHC TASSILI 21MDU323.

Hence, to ensure an accurate and robust sound recognition, a noise reduction method is required.

Various methods have been developed to solve the problem of interfering noises. Brown et al. (2017) recommended the Minimum Mean Square Error Short-time Spectral Amplitude Estimator (*MMSE-STSA*) for birds recordings. More recently, Deep learning-based noise reduction showing remarkable success by Fang et al. (2020). Nevertheless, this demands an immense dataset for training and is too complicated to be applied in real-time processing. Because of its low computational complexity, others like Weerasena et al. (2018) implemented a spectral subtraction in a sensor node inspired by Boll (1979). Nevertheless, it suffers from musical noise. In a constrained environment such as sensor networks, the complexity of noise reduction method must be considered to select a suitable robust and simple denoising method. In view of all that has been mentioned so far, one may suppose that not all types of denoising algorithms are relevant for our case as most of them do not consider energy efficiency. So far, and to the best of our knowledge, this is the first work that study an improved method of spectral subtraction, the Geometric Approach *GA* proposed by Lu and Loizou (2008), for birds call denoising. Also, very few works have attempted to integrate audio data in a sensor network for energy-efficient purposes.

Through this paper, we propose an accurate and energy-efficient system to monitor endangered birds in complex environment. We focus on the impact of the chosen audio-denoising technique on the classification performance. Also, we examine the energy consumption of our proposed overall system for reducing the processing and transmission of visual data in WMSN. The remainder of the paper is organized as follows. Section 2 describes our proposed monitoring system. The performance evaluation of the overall system is given in Section 3. Finally, the conclusion and perspectives are outlined in Section 4.

## 2. THE PROPOSED SYSTEM

Our proposed system, is devoted to minimize the energy consumption per node for visual data processing and transmission to prolong the network lifetime (Fig. 1). This can be provided by identifying the captured audio data in a sensor node before triggering the camera. The audio data is first processed, denoised and then classified as target species or not. If a target species is identified (including false positive ones), an image is captured, otherwise the camera stays on standby. Prior to its transmission to the end user, the image is compressed using a low cost encoding technique.

The proposed real-time audio recognition phase comprises three main steps. The first step consists in denoising the sensed audio signal to eliminate environment noises. To do so, we made use of a new geometric approach (*GA*) to spectral subtraction proposed by Lu and Loizou (2008) that improves the traditional spectral subtraction used by Weiss et al. (1975). Despite the fact that the latter is less complex, the former solves known shortcomings of the former like the musical noise and signal distortion.

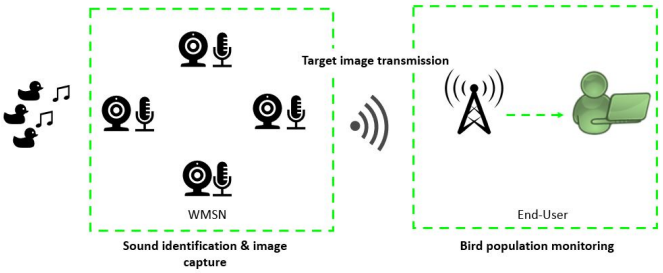


Fig. 1. WMSN for birds surveillance

Then, to only consider meaningful information of the audio signal, distinctive features are extracted using Mel Frequency Cepstral Coefficients (*MFCC*), key element of audio recognition. Their role is to identify the relevant components of the audio signal by eliminating irrelevant information. They are characterized by their computational efficiency and high performance under clean conditions Gupta et al. (2013).

Finally, the extracted features are compared to reference ones using Dynamic Time warping (*DTW*) distance for classification purposes using a threshold based matching. DTW is one of the known low-complexity algorithms for measuring the similarity between two time sequences. It aligns two feature vector sequences by repeatedly warping the time axis to find the optimal match between the two vectors Müller (2007). If the processed sound is identified as a target, the camera will be activated and an image of the target is captured to be transmitted to the end user.

The decision of capturing images is based on the results of audio classification. Visual data is compressed at the sensor node before being sent to the end-user. We applied a low complexity image compression using the open source tool Sensevid Maimour (2018).

## 3. PERFORMANCE ANALYSIS OF THE PROPOSED SYSTEM

In this section, we examine the proposed overall system. Firstly, we study the performance in terms of precision to highlight the impact of the the chosen pre-processing step (Denoising) on the classification accuracy. Secondly, we appraise the usage of energy by the overall system. We assume that the monitoring is performed periodically $N$ times during a day and compare three diverse scenarios (Figure 2) :

- $S0$: an image is captured and transmitted periodically with no local audio identification ;
- $S1$: an image is only captured and transmitted when a local audio identification detects a bird of interest excluding the denoising phase ;
- $S2$ (our proposed system) that operates as in scenario $S1$ but with a denoising phase prior to the audio identification.

### *3.1 Comparative Study of Denoising Methods*

We first carried out a performance comparison of *GA* against that of *MMSE-STSA* Brown et al. (2017) in terms of time execution and noise removal efficiency.

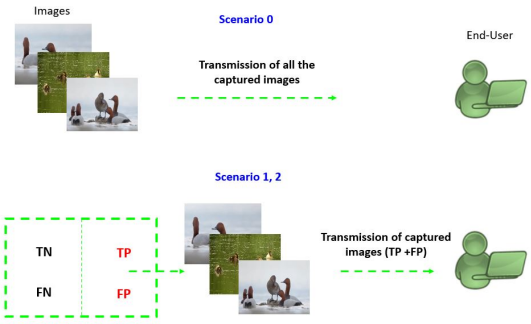


Fig. 2. Scenarios

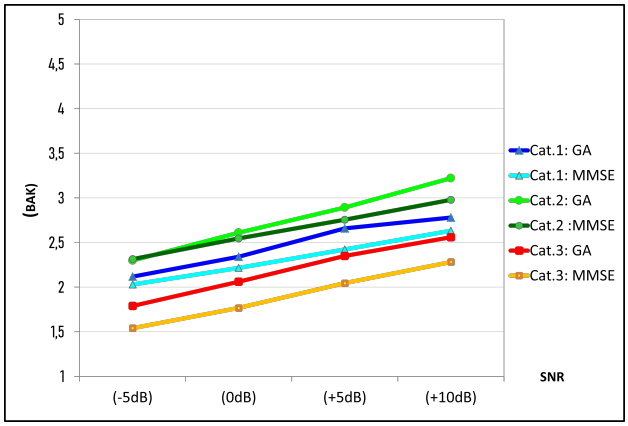


Fig. 3. BAK-based denoising evaluation

We are interested in studying three different species of wetland birds threatened of extinction. Table 1 summarises their types and characteristics. Since we cannot find bird's noisy songs dataset that covers diverse cases of noises interference encountered in nature, we created a dataset that contains about 500 noisy calls of the target specimens. The calls were manually polluted with different types of noises presented in wetlands at different intensities. Noises were chosen based on our multiple visits to the natural habitat during all seasons of the year. The dataset was created by the help of AUDACITY [1] tool.

We considered three categories of target birds, where each category contains 12 noisy sounds for testing. The calls were corrupted by three types of noise (wind, rain with thunder, and lake atmosphere). Four values of signal to noise ratio (SNR) were considered (-5dB, 0dB, +5dB and 10dB). The average duration of each recording is 2 seconds.

The denoising is a pre-processing step in the overall automatic sound-recognition system. Therefore, it should not take too long for being processed. Otherwise, the majority of the sensor energy will be wasted. We found that *GA* is about 9 times faster than *MMSE-STSA* for the three different categories of birds. Therefore, *GA* appears better suited for our needs. Despite that, we also evaluated it in terms of signal quality performance to approve our choice.

Since collecting bioacoustic data has increased, new efficient evaluation metric was created and has become an important metric for building real-time processing systems, so-called composite measures Hu and Loizou (2006). These measures were specially intended to evaluate audio enhancement techniques by predicting the quality of the enhanced noisy signal. They are derived by linearly combining existing objective measures involving subjective evaluations. In this work, we present our results based on the Background Noise Distortion (BAK) measure having five-point scale which varies from 1 (very conspicuous) to 5 (not noticeable). Figure 3 shows that *GA* outperforms *MMSE-STSA* for the three categories for removing noise. The present study confirmed that using a low complexity method is worthy since it has a very acceptable performance.

[1] https://www.audacityteam.org/

### *3.2 Impact of Denoising on the Classification Accuracy*

Here, the principal focus is to investigate the impact of the denoising step on the classification performance. In this part, we selected from our constructed dataset, 117 noisy target calls corrupted by the three different environmental noises at three different SNR measurements : 0 dB , 5 dB and 10 dB. In order to assess the behavior of the system when faced with non-target sounds, we added another 50 recordings. Among them, 30 are retrieved from Xeno-Canto website [2] and belong to 11 non-target bird species that live in the same area as our three target species. The remaining 20 recordings concern other various sounds (insects, water, other animals ...etc.) that can be encountered near lakes, the natural habitat of our target birds. We, finally, used 40 clean target audio samples for reference. In total, 167 recordings were tested.

To attest the performance of our classification, the following metrics were considered :

$$Precision = TP/(TP + FP)$$
$$Recall = TP/(TP + FN)$$
$$Fmeasure = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$
$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

where $TP$, $TN$, $FP$ and $FN$ represent true positive, true negative, false positive, and false negative respectively. Precision is the proportion of relevant audio data among the retrieved sounds in the dataset. Recall (or sensitivity) indicates the ratio of the number of correctly identified birds to the total number of relevant calls in the dataset. $F$ measure balances between precision and sensitivity. Accuracy denotes the proportion of the total number of predictions that were correct in the dataset. The matrix confusion is represented in Table 2.

Figure 4 provides the classification evaluation of the audio chain system in $S1$ and $S2$ respectively. We can clearly notice that the results of $S2$ are statistically significant for all metrics compared to that of $S1$. The low performance of $S1$ for Recall, F-measure and Accuracy is attributable to the distortion of the audio call. This last was corrupted by environmental noises, thus, the system was not able to identify the majority of the target bird calls. However,

[2] `https://www.xeno-canto.org/`

Table 1. Target Birds Categories and Noises

| Category | Target species | Scientific name | Noise type |
|---|---|---|---|
| Cat-1 | Ferruginous Duck | Aythya nyroca | Wind |
| Cat-2 | Common Pochard | Aythya ferina | Rain with thunder |
| Cat-3 | White Headed Duck | Oxyura leucocephala | Lake Atmosphere |

Table 2. Matrix confusion

| | TP | FP | FN | TN |
|---|---|---|---|---|
| *S1* | 13 | 2 | 104 | 48 |
| *S2* | 93 | 2 | 24 | 48 |

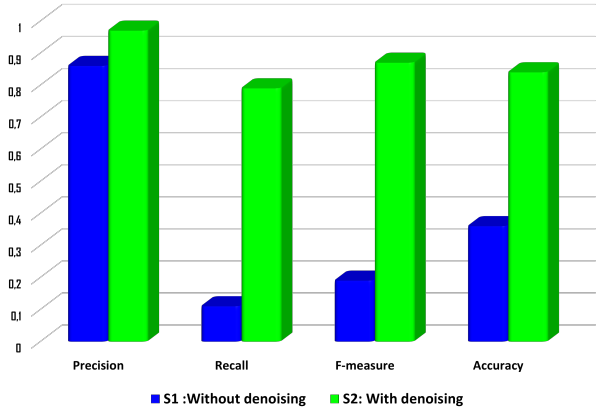


Fig. 4. Classification performance

it was able to identify the non-target calls much better due to the threshold used for classification. This is what justify the high value of Precision of $S1$. In fact, both $S1$ and $S2$ were able to identify the non target calls with equal performance. The only difference is that in $S2$, we recognized birds call much better. In summary, according to all performance metrics, the chain audio in $S2$ achieved better performance than in $S1$, confirming the importance role of the denoising phase in our proposal.

### *3.3 Energy Consumption*

The main focus of this section is to investigate and estimate the energy consumption per day from end-to-end for our different scenarios that includes the amount of energy needed to process an audio recording $E_{audio}$ and the energy required to capture, encode and transmit one image $E_{img}$. The energy can be computed for a given processor using :

$$E = processor\_power \times cycles/clock\_rate \quad (1)$$

where *cycles* denotes the number of cycles required to perform a given processing. The total power of the audio identification system can be computed using :

$$P_{audio} = P_{op} \times \sum_{op} \sum_{i} N_{op}^{i} \quad (2)$$

where $N_{op}^{i}$ is the number of operations $op$ (multiplication, division, addition, subtraction) required to perform a given phase $i \in \{GA, MFCC, DTW\}$ and $P_{op}$ is the power consumption of operation $op$. Table 3 gives for one audio-frame, the number of operations required to perform the different stages of the audio recognition model.

With respect to our three scenarios, the total required energy consumed by the network is given by :

Table 3. Computational cost for one audio-frame

| Step | Mult. | Div. | Add. | Sub. |
|---|---|---|---|---|
| *Denoising (GA)* | 14353 | 7176 | 3601 | 7176 |
| *Feature extraction (MFCC)* | 2485 | 1636 | 1073 | 2218 |
| *Classification (DTW)* | – | – | 11 | 3 |

Table 4. Cortex M3 parameters

| | |
|---|---|
| Sensor Processor | Cortex M3 |
| Clock rate | 72 MHz; |
| Processor power | 23 mW; |
| Cycles count | Add.[1], Sub.[1], Mult.[1 or 2], Div.[1 to 12]. |

Table 5. Energy consumption of the different phases.

| Step | Energy consumption ($mJ$) | |
|---|---|---|
| | Min. | Max. |
| Denoising ($GA$) | 0.59 | 2.29 |
| $MFCCs$ extraction | 0.29 | 1.13 |
| $DTW$ classification | 0.10 | 0.10 |
| Image capture | 9.540 | 9.540 |
| Image encoding | 8.058 | 8.058 |

$$E = \begin{cases} N\ E_{img} & \text{for } S0 \\ (TP + FP)E_{img} + N\ E_{audio} & \text{for } S1 \text{ and } S2 \end{cases} \quad (3)$$

In what follows, we consider a sensor node equipped by an ARM Cortex M3 micro-controller cor (2018) and an ATMEL radio interface. Its main features are summarized in Table 4. Table 5 shows the energy consumption for each step of the overall system including image energy demand derived using SenseVid for an image of resolution $640 \times 360$. The estimated maximum and minimum energy for the audio chain are given for a 2-second audio call and is related to the maximum and minimum number of cycles for multiplication and division operations. It can be derived that the total energy required for visual data is much higher than for audio data. All the other used parameters in our study are provided in Table 6.

Figure 5 plots the energy consumed in the different scenarios as a function of the probability $p$ that a target bird occurs in one period during the day. It is assumed that the sensor node is able to reach the collect station in one hop. For $S1$ and $S2$, we have estimated the maximum and minimum of energy according to the energy demand of audio chain system. It is clear that $S0$ consumes more energy than the other two scenarios. Its energy consumption remains stable since it does not depend on TP and FP values. We send all the captured images, including the ones of interest and non of interest. Also the higher consumption of $S0$ refers to the total energy consumed by each image. The low energy consumption of $S1$ compared to $S2$ is explained by the low number of TP. The system was not able to identify well the target birds, as a consequence, fewer images were processed.

Table 6. Experimental Setup

| | |
|---|---|
| Sound duration | 2 seconds |
| Frame samples | 444 |
| Sampling frequency | 22050 Hz |
| Coding | 16 bits |
| | |
| Image resolution | 640x360 |
| Size compressed image | 4863 Bytes |
| Packet maximum payload | 128 Bytes |
| Number of packets | 50 |
| DCT | Triangular BIN DCT; $\rho$=8 |
| Entropy coder | Exponential-Golomb (EG) |
| | |
| Sensor Data Rate | 250 kBytes/sec |
| Number of hops | 1, 2 and 7 |
| Reception sensitivity | -101 dBm |
| $I_{Rx}$ | 12.3 mA |
| $I_{Tx}$ | 14mA; 11.6mA; 7.4mA. |

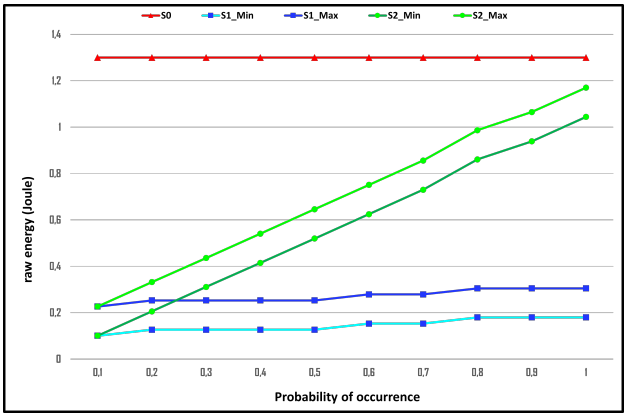


Fig. 5. Overall consumed energy

Figure 6 represents the energy consumed per byte estimated by dividing the overall energy by the amount of visual data transmitted :

$$E_B = \begin{cases} \dfrac{E}{N \times imageSize} & \text{for } S0 \\ \dfrac{E}{(TP+FP) \times imageSize} & \text{for } S1 \text{ and } S2 \end{cases} \quad (4)$$

Note that this counts for all transmitted images (after compression) including target and non target ones. For both *scenarios* 1 and 2, the number of transmitted images is the sum of TP and FP. The highest values of ratio belong to $S1$, as we transmit less images and we loose significant amount of important data. In contrast to $S2$ where we send a satisfactory amount of data of interest even though we loose some. For $S0$, it has the lowest ratio since we send the total of captured images without selection.

Figure 7 plots the useful energy per byte given as follows :

$$E_U = \begin{cases} \dfrac{E}{N \times imageSize} & \text{for } S0 \\ \dfrac{E}{TP \times imageSize} & \text{for } S1 \text{ and } S2 \end{cases} \quad (5)$$

For both scenarios 1 and 2, the number of transmitted images of interest is the number of TP. It is clear that $S2$ has the minimum ratio, because the majority of transmitted visual data are of interest, which is not the case for $S0$ and $S1$.

Figure 8 illustrates the wasted energy estimated by the following equation :

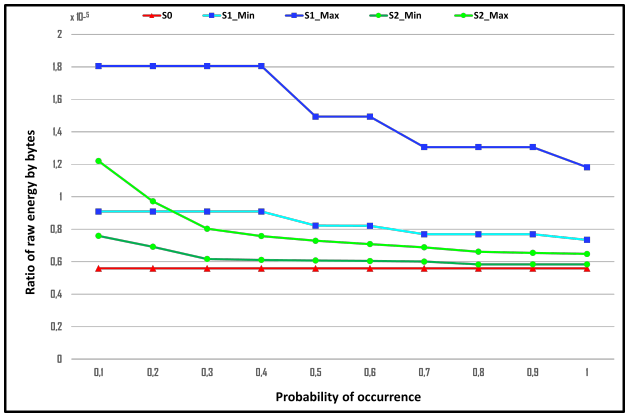


Fig. 6. Energy per byte

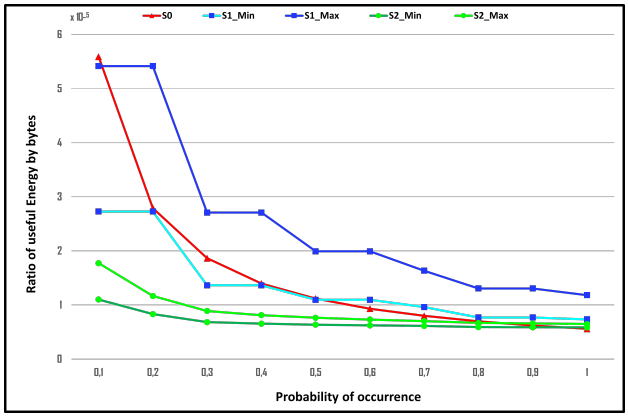


Fig. 7. Useful energy

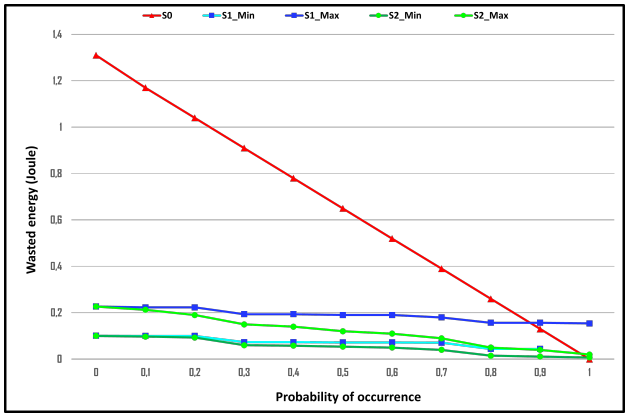


Fig. 8. Wasted energy

$$E_W = \begin{cases} (FP+TN+FN)E_{img} & \text{for } S0 \\ FP \times E_{img} + (FP+TN+FN)E_{audio} & \text{for } S1 \text{ and } S2 \end{cases} \quad (6)$$

We note that $S0$ wastes about 1.3 J when the probability of occurrence is zero, which is about 6 times higher than $S1$ and $S2$. Then it decreases with the probability of the occurrence of an event of interest. The plots also show that $S2$ wastes less energy than $S1$, which is due to the low FN values compared to those in $S1$.

Figure 9 demonstrates a comparison of useful energy consumption in the three scenarios when the number of hops toward the collect station is raised to 2 and 7. We averaged the maximum and minimum values and set the

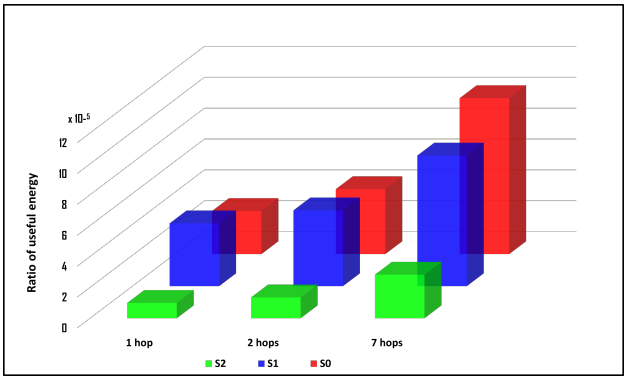


Fig. 9. Useful energy for 1, 2 and 7 hops

probability of occurrence of target birds to 0.2. We can see that the energy consumption for 7 hops is higher than for 2 hops, and that the latter consumes more energy than 1 hop. This is attributed to the required energy for transmission and reception of images in each node. $S2$ seems to be the preferable compared to the other scenario for the three number of hops in the network.

As a conclusion, the findings attest that $S2$ or our proposed system is the most suitable scenario. It does not waste excessive energy as in $S0$, which consequently prolongs the energy gain of the node's lifetime and improves the network performance. Furthermore, it does not loose important informative data as in $S1$, which is the best compromise between preserving energy and interest data at the same time.

## 4. CONCLUSION AND PERSPECTIVES

This paper presents a proposed method to extend the lifetime of WMSN for our threatened bird's surveillance application. The idea is to reduce unnecessary visual data processing and transmission through low-energy local identification based on audio bird calls. The detection of the target species in the area will trigger the camera connected to the sensor node. However, in a complex acoustic environment, a denoising method is required for robust and improved automatic bird call detection.

In this work, we have focused on evaluating the performance of the proposed system in terms of accuracy and mainly energy consumption. The experimental results show that the proposed method performs well in identifying bird calls thanks to the denoising phase. Furthermore, our test results attest that the proposed system is able to achieve a per-node energy gain and therefore extend its lifetime.

In future work, we will investigate the impact of an adaptive threshold on the accuracy of audio bird classification. Furthermore, we will optimize the audio chain, in particular the denoising phase, with the goal of minimizing the energy consumption per-node.

## REFERENCES

(2018). Cortex m3 datasheet. URL `iot-lab.github.io/assets/misc/docs/iot-lab-m3/stm32f103re.pdf`.

Allinson, T. (2018). State of the world's birds. Technical report, BirdLife International.

Almalkawi, I.T., Guerrero Zapata, M., Al-Karaki, J.N., and Morillo-Pozo, J. (2010). Wireless multimedia sensor networks: current trends and future directions. *Sensors*, 10(7), 6662–6717.

Boll, S. (1979). Suppression of acoustic noise in speech using spectral subtraction. *IEEE Transactions on acoustics, speech, and signal processing*, 27(2), 113–120.

Brown, A., Garg, S., and Montgomery, J. (2017). Automatic and efficient denoising of bioacoustics recordings using MMSE STSA. *IEEE Access*, 6, 5010–5022.

Civelek, M. and Yazici, A. (2016). Automated moving object classification in wireless multimedia sensor networks. *IEEE Sensors Journal*, 17(4), 1116–1131.

Fang, H., He, Y., Xu, W., Xu, Y., Ke, D., and Su, K. (2020). A generalized denoising method with an optimized loss function for automated bird sound recognition. In *2020 5th International Conference on Computer and Communication Systems (ICCCS)*, 240–245. IEEE.

Gupta, S., Jaafar, J., Ahmad, W.W., and Bansal, A. (2013). Feature extraction using mfcc. *Signal & Image Processing: An International Journal*, 4(4), 101–108.

Hu, Y. and Loizou, P.C. (2006). Evaluation of objective measures for speech enhancement. In *Ninth international conference on spoken language processing*.

Koyuncu, M., Yazici, A., Civelek, M., Cosar, A., and Sert, M. (2018). Visual and auditory data fusion for energy-efficient and improved object recognition in wireless multimedia sensor networks. *IEEE Sensors Journal*, 19(5), 1839–1849.

Lu, Y. and Loizou, P.C. (2008). A geometric approach to spectral subtraction. *Speech communication*, 50(6), 453–466.

Magno, M., Tombari, F., Brunelli, D., Di Stefano, L., and Benini, L. (2013). Multimodal video analysis on self-powered resource-limited wireless smart camera. *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, 3(2), 223–235.

Maimour, M. (2018). SenseVid: A traffic trace based tool for QoE video transmission assessment dedicated to wireless video sensor networks. *Simulation Modelling Practice and Theory*, 87, 120–137. URL `https://www.sciencedirect.com/science/article/pii/S1569190X1830090X`.

Müller, M. (2007). Dynamic time warping. 69–84. Springer, Berlin, Heidelberg. URL `https://doi.org/10.1007/978-3-540-74048-3_4`.

Weerasena, H., Jayawardhana, M., Egodage, D., Fernando, H., Sooriyaarachchi, S., Gamage, C., and Kottege, N. (2018). Continuous automatic bioacoustics monitoring of bird calls with local processing on node level. In *TENCON 2018-2018 IEEE Region 10 Conference*, 0235–0239. IEEE.

Weiss, M.R., Aschkenasy, E., and Parsons, T.W. (1975). Study and development of the intel technique for improving speech intelligibility. Technical report, NICOLET SCIENTIFIC CORP NORTHVALE NJ.

Xie, J., Colonna, J.G., and Zhang, J. (2021). Bioacoustic signal denoising: a review. *Artificial Intelligence Review*, 54(5), 3575–3597.

ZainEldin, H., Elhosseini, M.A., and Ali, H.A. (2015). Image compression algorithms in wireless multimedia sensor networks: A survey. *Ain Shams engineering journal*, 6(2), 481–490.