Dynamic sensor activation and decision-level fusion in Wireless Acoustic Sensor Networks for classification of domestic activities

Gert Dekkers[a,b,*], Fernando Rosas[c,d,e], Toon van Waterschoot[b], Bart Vanrumste[b], Peter Karsmakers[a]

[a] *KU Leuven, Department of Computer Science, Geel, Belgium*
[b] *KU Leuven, Department of Electrical Engineering, Leuven, Belgium*
[c] *Data Science Institute, Imperial College London, UK*
[d] *Department of Brain Sciences, Imperial College London, UK*
[e] *Centre for Complexity Science, Imperial College London, UK*

Abstract

For the past decades there has been a rising interest for wireless sensor networks to obtain information about an environment. One interesting modality is that of audio, as it is highly informative for numerous applications including automatically classifying domestic activities that is focussed on in this work. However, as they operate at prohibitively high energy consumption, commercialisation of battery-powered wireless acoustic sensor networks has been limited. To increase the network's lifetime, this paper explores decision-level fusion, adopting a topology where processing – including feature extraction and classification – is performed on a (dynamic) set of sensor nodes that compute classification outputs which are fused centrally. The main contribution of this paper is the comparison of decision-level fusion with different dynamic sensor activation strategies that leverage the redundancy of information in the network. Our results show that representing the classification output using vector quantisation can reduce communication per classification output to 8 bit without loss of significant performance. In case of fixed sensor activation this results in an energy reduction up to 3%. While the savings of fixed sensor activation are limited, it is shown that dynamic sensor activation, using a centralised approach, can provide an energy reduction up to 80%. In general, this work indicates that if opted for a topology using decision-level fusion, dynamic sensor activation is needed when a long battery lifetime is desired.

Keywords: Sound classification, Activities of the Daily Living, Wireless Acoustic Sensor Network, Edge computing, Decision-level fusion, Dynamic sensor activation

1. Introduction

Over the past decades, integrated components containing wireless radios and sensors are shrinking in size, while maintaining computational power [1]. This has facilitated the rising

*Corresponding author

Email addresses: gert.dekkers@kuleuven.be (Gert Dekkers), f.rosas@imperial.ac.uk (Fernando Rosas), toon.vanwaterschoot@esat.kuleuven.be (Toon van Waterschoot), bart.vanrumste@kuleuven.be (Bart Vanrumste), peter.karsmakers@kuleuven.be (Peter Karsmakers)

interest in smart environments, which aim to understand the home scene for enabling smart functionality to the inhabitant, e.g. security, health monitoring and entertainment [2, 3, 4, 5, 6]. The use case depicted in this paper is the classification of domestic activities, including the Activities of the Daily Living (ADL), which describes the current activity being performed by a person in a domestic environment (e.g. cooking or watching TV).

A first application to automatic classification of ADLs is that it plays a vital part in a system to measure the self-reliance of a person. The ratio of retired to working people is significantly increasing, which brings important challenges to our society. One of the main wishes of the ageing population is to be able to live in their own dwelling as long as possible. Self-reliance is currently manually determined by healthcare professionals using validated scales. Observing the self-reliance continuously and automatically individualises the care of older persons [7].

Another application is related to comfort, where such information could be fed to a domotics system to automatically control an actuator (e.g. activate extraction hood when cooking, dim lights when watching TV). Similarly, it can also provide information to a security system (e.g. when the owner is not present no activity should be detected).

To ensure sufficient spatial coverage for classifying domestic activities, wireless sensor networks are of interest [8, 9]. In this work microphones are used as sensor, thereby making it a Wireless Acoustic Sensor Network (WASN). A microphone contains highly informative data which can be leveraged for classifying domestic activities [9] but also for other tasks, e.g. speech recognition, urban scene classification and sound event detection [10, 11, 12, 13]. In the remainder of this paper a *sensor node* refers to a single device in a WASN capable of sensing, processing and communication.

In order to make such a system easily installed, an interesting avenue is to use battery-powered sensor nodes in the WASN [8]. This brings an additional challenge because, besides acquiring a high performance, a long autonomous lifetime of the sensor nodes is of interest. Reducing energy consumption of a particular sensor node can be tackled on many layers of the processing chain (e.g. sensing, feature extraction, classification and communication) [14]. An important choice to make before designing a WASN is to determine how much processing is performed locally, i.e. on the sensor nodes instead of in a fusion center. Performing more computation locally has the potential advantage of a lower communication bandwidth at the expense of less information being available centrally and vice-versa.

The first part of the paper explains the motivation for, and evaluation of, different decision-level fusion methods. Each sensor node in a network performs local processing (including classification), and these local classification outputs (decisions) are combined centrally in a fusion center to obtain a final classification output based on local decisions. The evaluated methods differ on the needed communication bandwidth and are therefor evaluated on classification performance at the fusion center along with the communication bandwidth and energy consumption that is needed to compute a single classification output at the local level.

The second part of the paper introduces and evaluates dynamic sensor activation algorithms. As the sensor nodes in the network monitor the same process from a different point of view, redundancy may exist and a *good* subset of the sensor nodes may be sufficient. The proposed algorithms dynamically (de-)activate layers of the processing chain (i.e. sensing, processing and communication) for each sensor node separately to further reduce energy consumption.

In this work, related to classification of domestic activities, the main contributions are:

(a) A motivation for local classification based on an energy consumption model,

(b) A performance baseline for single-sensor node versus multi-sensor node classification,

(c) An evaluation of the classification performance and communicated bits for different decision-level fusion methods with a fixed set of active sensor nodes,

(d) A comparison of three dynamic sensor activation strategies using decision-level fusion on classification performance and the resulting percentage of time the sensor node is active (i.e. duty cycle).

The rest of the paper is organised as follows. In Section 2, the related work is introduced. As this paper is cross-disciplinary this section introduces work from all involved disciplines. Section 3 introduces the problem description along with a motivation for the design choices that were made. In order to properly compare the proposed algorithms two data sets are used which are introduced in Section 4. Section 5 shows the results for single sensor node classification. Section 6 and 7 introduce and discuss the results related to the proposed algorithms. A final discussion on the results along with conclusions are provided in Section 8.

2. Related work

First, the literature regarding sound monitoring in general and specifically for classifying activities in domestic environments is briefly discussed. Then, relevant work regarding data fusion is reviewed. Finally, approaches available in the literature to make WSN more energy-efficient are summarised.

Audio is an attractive sensor modality for monitoring an environment as it can convey highly informative data [10, 7, 15]. Over the past decade, a considerable amount of research has been performed on designing computational methods to automatically extract information from audio data. Besides the classification of domestic activities, depicted in this paper, many other tasks are researched [11, 12, 13]. Few examples are, but not limited to, classification of urban acoustic scenes [16], general-purpose audio tagging [17], bird sound detection [18] and sound event detection in domestic environments [19]. In many of these tasks similar techniques are being used, starting from a paradigm consisting of a hand-crafted feature extraction and a machine learning stage. In recent years, (Deep) Neural Networks have become the most popular computational method, which has led to a paradigm shift where feature extraction is no longer or less hand-crafted and included in the learner's objective. An overview of the current state-of-the-art for various sound recognition tasks can be found in [11, 12, 13].

Regarding the task depicted in this paper, classification of activities in a domestic environment, research has been devoted to a wide range of sensor modalities [7]. Research using solely audio or a combination of sensor modalities is limited. In [10, 20] a system is introduced for ADL recognition along with distant speech recognition for home automation. The dataset in that work contained multiple non-wearable sensor modalities including audio. Regarding audio, the number of audio events was extracted using an adaptive threshold algorithm and used as a feature. This is similar to the work performed in [7] which focusses solely on classification of ADLs using a multi-modal dataset. In [21], the authors analyse the performance of a system with Mel-Frequency Cepstral Coefficients (MFCC) feature extraction along with a Support Vector Machine or Gaussian Mixture Model classifier with respect to computational complexity for various parameter settings. Recently, a competition was organised in the scope of the DCASE 2018 Challenge, which was related to classification of ADLs using multi-channel audio [22]. Most submissions used log-mel energies or MFCCs as features along with a Neural Network-based classifier.

In case multiple sensor nodes are available, information needs to be fused to output a final decision. By fusing information, the system can increase its spatial resolution and consequently achieve a higher performance [8]. Multi-sensor data fusion is widely used to combine data acquired by different sensor nodes to monitor a particular process [23, 24, 25, 26]. Typical approaches can be subdivided into data-, feature- and decision-level fusion and whether or not labels are available. Fusing data at an earlier stage, either as raw data or as features that are a compressed representation of the raw data, can be beneficial to obtain a better classification performance, while fusing at a later stage, i.e. decisions, could reduce communication bandwidth. Unfortunately, the problem of finding optimal strategies for distributed processing and decision making is in general NP-hard [27]. Therefore, heuristics (see e.g. [28]) are necessary when designing practical solutions.

Decision-level fusion approaches, such as the ones considered in this paper, can be categorised based on how the output of each classifier is formatted (e.g. labels or probability values) [29]. Common methods that do not require a learning cycle that needs data annotation are the product, sum, maximum and minimum rule in case of probabilistic inputs and the majority vote or borda count in case of labels [24, 29, 30]. To improve the performance of decision-level fusion researchers have suggested (optimal) fusion strategies that require learning or a-priori information [24, 26, 29, 31]. In this paper methods are explored that do not need a learning cycle. Methods which do require training need labelling and training for each environment, which is not a realistic setting as positions of the sensor nodes and the deployed environments are different for each setup. As a consequence, for each environment, a new training cycle would be required.

The literature on sound classification tailored to WASN is limited. In [32] the authors propose a WASN for classifying ADLs. The benefit of a WASN, by using a Signal-to-Noise ratio (SNR) based sensor selection, is shown in clean and noisy conditions. In [33] various decision-level fusion methods are compared for acoustic event detection and classification using a Bag-of-Features type of classification. Similarly, in [34] multiple decision-level fusion methods are compared for audio event detection on different SNRs. Different from their work is that here not only performance but also the energy consumption is taken into account.

Increasing the energy efficiency of a WASN can be tackled in different layers of the processing chain, including sensing, signal processing and wireless communication [14, 35]. With respect to the processing layer, reduction of energy consumption can be achieved by designing energy-efficient hardware [36, 37], or comparing different algorithms and parameter settings [21, 38]. Substantial efforts have been made to decrease energy consumption of wireless communication modules, ranging from the physical layer [39, 40, 41, 42], multihop and routing [43, 44, 45] to network layer protocols [46]. Another approach is to consider different communication strategies as in [15] where the energy consumption for acoustic surveillance was evaluated on a distributed and centralised approach for sound source detection and localisation. A characteristic of a WASN is that the collected data and obtained information at a sensor node is correlated to that of neighbouring sensor nodes, which could be used to reduce the overall energy consumption [47]. To the best of our knowledge no work has been performed on comparing (dynamic) decision-level fusion schemes with respect to performance and communication bandwidth for the purpose of audio classification.

3. Problem description and motivation

This work investigates the distribution of processing tasks over different components of a WASN with battery-fed sensor nodes in order to optimise the autonomy of the nodes. A specific

use case of automated monitoring of domestic activities of persons was selected to carry out the study. For this use case the task of the WASN is to classify audio segments into a pre-defined set C of N_C daily activities. First, the WASN setup that is used throughout the paper is discussed. Then, a motivation is provided for using decision-level fusion. Finally, the problem definition for dynamic sensor activation is introduced.

3.1. WASN setup and energy model

In Figure 1 a WASN is shown that consists out of multiple acoustic sensor nodes with wireless communication capabilities and a central dedicated device or *fusion center*, that can gather and process (fuse) the sensed data. In order to allow such a WASN to be easily installed, wireless battery powered architectures are preferred to avoid extensive use of wiring [8]. Unfortunately, this brings additional challenges as the lifetime of these devices can be compromised by the energy consumption of acoustic sensors $\mathcal{E}_\mathrm{S}$, local processing $\mathcal{E}_\mathrm{P}$, and wireless transmission $\mathcal{E}_\mathrm{T}$ and reception $\mathcal{E}_\mathrm{R}$, which usually goes beyond the scope of what current sensor network architectures can provide [48]. Note that the following characteristics of the WASN were assumed: star topology, sensor nodes support duty cycling, an up- and downlink is available and the fusion center is not energy-constrained, while the others are battery-powered.

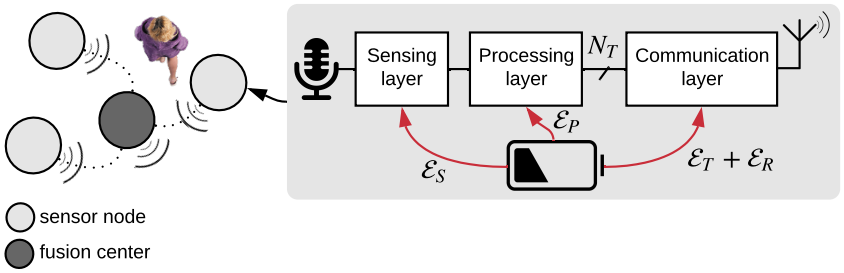


Figure 1: A WASN containing a fusion center along with multiple sensor nodes. Each sensor node is represented in three basic layers along with their energy consumption, i.e. sensing, processing and communication layer with $\mathcal{E}_\mathrm{S}$, $\mathcal{E}_\mathrm{P}$ and $\mathcal{E}_\mathrm{T} + \mathcal{E}_\mathrm{R}$ respectively. [49]

To motivate the experiment conducted in this paper, an energy consumption model is introduced. The total energy cost of a sensor node can be calculated as:

$$\mathcal{E}_\mathrm{node} = \delta_S \mathcal{E}_\mathrm{S} + \delta_P \mathcal{E}_\mathrm{P} + \delta_T \mathcal{E}_\mathrm{T} + \delta_R \mathcal{E}_\mathrm{R} \tag{1}$$

which is comprised of the energy consumption of acoustic sensors $\mathcal{E}_\mathrm{S}$, local processing $\mathcal{E}_\mathrm{P}$, and wireless transmission $\mathcal{E}_\mathrm{T}$ and reception $\mathcal{E}_\mathrm{R}$. Reducing the duty cycle $\delta \in [0, 1]$ has an impact on a specific layer. The respective duty cycle for a particular layer is denoted with the respective subscript. In this work, duty cycle refers to a percentage of the average time a layer (i.e. sensing, processing and communication) is active compared to always-on. Note that the energy cost for reception $\mathcal{E}_\mathrm{R}$ is only needed if communication is needed from the fusion center to a dedicated node. For simplicity, our modelling neglects costs associated with feedback control messages, which is often orders of magnitude smaller than the costs of feedforward transmissions [50, 51, 52].

3.2. Motivation for choosing decision-level fusion

Reducing the amount of bits to communicate from a sensor node to the fusion center involves processing of the collected data which increases the energy consumption, on the sensor node, due to arithmetic operations, memory storage and memory accesses. To gain autonomous lifetime, the additional energy consumption should -obviously- be less than the energy consumption decrease in the communication layer.

Figure 2 shows the energy consumption of each component in a hypothetical WASN in function of the amount of transmitted bytes per classification output. To classify the acoustic data, it is assumed that a segment of 15s of data is available such as in [9]. The sensing layer, contributing to cost $\mathcal{E}_{\mathrm{S}}$, outputs an amount of bits depending on the bit depth S and sampling frequency f_s of the analog-to-digital converter. Subsequently, the processing layer processes the sampled audio to output features or a meaningful classification output (e.g. detected activity is cooking). In this figure the energy acquired by the processing layer is denoted as $\mathcal{E}_{\mathrm{P,low}}$ and $\mathcal{E}_{\mathrm{P,high}}$ referring to two architectures with a relatively low and high complexity respectively based on [9, 14]. The amount of communicated bytes, attributing to $\mathcal{E}_{\mathrm{T}}$, are roughly categorised in three groups or strategies for distributed classification: raw audio (24 kB/s with $f_s = 16$ kHz and $S = 12$ bit), feature extraction (in-between) and classification (160 bit per classification output in case of 10 classes with 16 bit posterior probabilities).

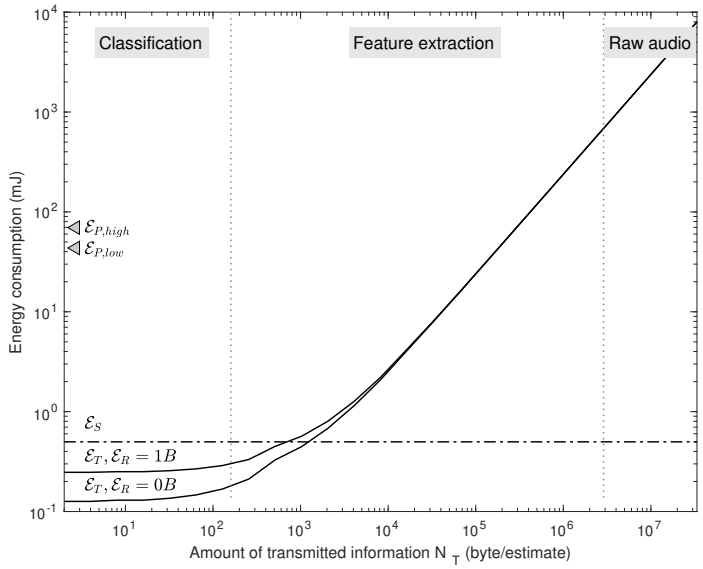


Figure 2: Energy consumption of each layer in a WASN in function of the amount of communicated bytes.[1]

[1]More information on the hypothetical hardware architecture and how the energy consumption is calculated can be found in [14]. The energy consumption model mimics the energy consumption behaviour of a setup including a microphone, analog-to-digital converter, microcontroller with a Cortex-M4 processor [53] and communication module following the IEEE 802.15.4 standard [54]. Open-source code to compute the graph is available at [49].

Regarding communication, two curves are shown, with and without additional cost $\mathcal{E}_{\mathrm{R}}$ for receiving information of a single byte from the fusion center. The cost of $\mathcal{E}_{\mathrm{R}}$ should only be taken into account if it is desirable to control each sensor node from the fusion center. Given this hypothetical sensor node and a typical battery of 2100 mAh at 2.2V, a lifetime is expected of ±20 and ±30 days for the low and high complexity architecture with sensing and communication (i.e. $\mathcal{E}_{\mathrm{P,low}}$ and $\mathcal{E}_{\mathrm{P,high}}$ respectively). In the case of communicating raw audio this would be ±2 days, i.e. not an option when a high autonomous lifetime is required. As a consequence the use of local processing that computes high-level features or classification outputs is desired. Evidently, the additional energy spent by local processing should remain limited to not undo the gain that is achieved by lowering communication bandwidth.

Previous paragraph indicated that lowering the communication bandwidth is desired if high autonomous lifetime is needed. In this work it is chosen to compare decision-level fusion methods w.r.t. communicated bits per classification output and classification performance including a fixed and dynamic sensor activation strategy.

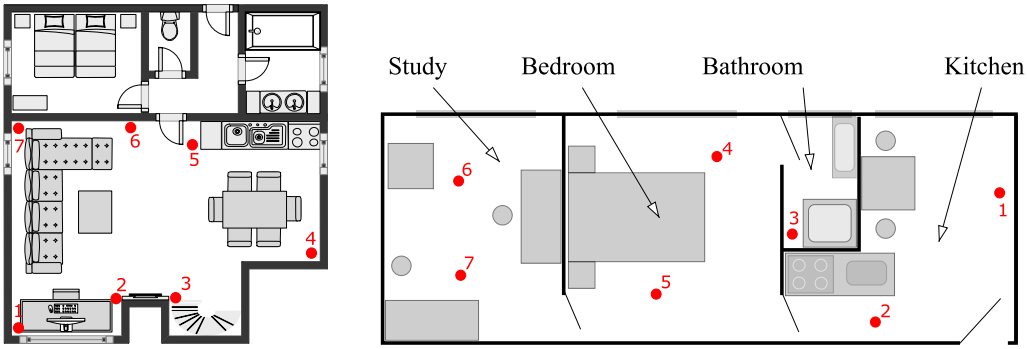
3.3. Dynamic sensor activation

With typical decision-level fusion, sensor activation is fixed. In this case, the processing layer tends to dominate to overall energy consumption. To further reduce energy consumption, dynamic sensor activation strategies are explored that can reduce the sensor node's duty cycle. When using a WASN, sensor nodes are spatially distributed and redundancy will exist. This was the motivation to evaluate algorithms that, based on the given classification outputs off one or multiple sensor node(s), select the next set of active sensor nodes. Three algorithms are proposed that differ on where the decision is made, i.e. locally, centrally or a hybrid form.

4. Datasets

The results in this work are based on two datasets containing multi-channel acoustic recordings of daily activities performed in a home environment. The first dataset, named *SINS*, contains audio of a single person living in a vacation home recorded using a WASN of seven sensor nodes each containing an array of four microphones. The daily activities are performed in a spontaneous manner and recorded in a continuous stream. In this work the data acquired by the first sensor in each of the seven microphone arrays in the living room is used. The annotation of the data acquired in the living room distinguishes ten activities: cooking, calling, dishwashing, eating, vacuum cleaning, visit, watching tv, working, absence and others. More information on the dataset can be retrieved in [9].

The second dataset, named *SWEET-HOME*, is a multi-modal dataset recorded in an apartment. In this work the audio recordings of the subset named *multimodal* are used. The subset contains recordings of 21 participants that perform several activities without any constraint on duration and procedure. The data was collected for each participant separately by seven sensor nodes containing a single microphone distributed over multiple rooms in the apartment. In total 26 hours of data is available per microphone channel. The dataset distinguishes between eight activities: cleaning, communication, dressing, hygiene, meal, relaxation, sleeping and other. Figure 3 shows a two-dimensional map of the recording environments of both datasets. Each map contains dots indicating the placement of each sensor node along with an index number.



(a) (b)

Figure 3: 2D map of the recording environment used in the (a) SINS and (b) SWEET-HOME [20] dataset

Both datasets are segmented such that each classification example is an audio sample of 15 seconds and each two successive samples overlap for 5 seconds, similarly as in [9]. Only a single class is active at a particular time instant. However, as these datasets are continuous, a single audio sample can contain a transition between two classes. Each audio sample is assigned a label corresponding to that of the class that is active in the middle of the sample. Table 1 shows the amount of audio samples per class for both datasets along with the amount of groups. A group represents a consecutive set of samples that belong to a single activity and should be kept together when dividing in training and test sets for classification. In total 106419 and 18819 audio samples are available for the *SINS* and *SWEET-HOME* dataset respectively.

Class	Samples	Groups
Absence	50977	64
Calling	2162	22
Cooking	3791	19
Dishwashing	1144	15
Eating	1771	19
Other	1789	198
Vacuum cleaner	745	13
Visit	1435	9
Watching TV	24238	13
Working	18367	49

(a)

Class	Samples	Groups
Cleaning	3048	60
Communication	664	43
Dressing	486	50
Hygiene	1317	60
Meal	3913	21
Other	3850	285
Relaxation	3689	81
Sleeping	1852	38

(b)

Table 1: Dataset distribution over all class labels for the SINS (a) and SWEET-HOME (b) dataset.

5. Classification using a single sensor node

5.1. Overview

Before exploring the performance of a WASN the limits in terms of classification accuracy and energy consumption of a single sensor node system are studied. The performance of a single

sensor node, for the given problem, varies and is subject to location, sensor node characteristics and classifier model variability. Therefore, for each available sensor first an individual performance evaluation was performed. Secondly, statistics are drawn from the performance of the entire sensor set, i.e. the best, average and worst sensor. In this section, and in the remainder of the paper, two classifier architectures of distinct computational complexity are used. Multiple architectures were tested and provided similar results, therefor only the two *extremes* are shown here.

5.2. Classifier architectures

For this experiment two classifier architectures were chosen with a relatively low and high computational complexity, denoted as *NN* and *CNN* respectively. These architectures are based on common architectures used in the field of sound classification including typical Mel-Frequency Cepstral Coefficients (MFCC) and a (Convolutional) Neural Network (NN) [11, 12, 13]. The computational complexity is intentionally kept relatively low such that it is realistic to fit on a low-power embedded device.

For both architectures, a Short-Time Fourier Transform (STFT) is applied to the audio samples of 15 s duration, that have a sampling frequency of 16 kHz. A 30 ms Hamming window and a 10 ms step size was used. Then, the STFT magnitude of the resulting frames is fed to a mel-scale filterbank with 26 bands and a frequency range of 500 to 8000 Hz followed by a logarithmic transformation at the end which leads to a feature representation of size 26 by 1500. Regarding the *CNN* architecture, the aforementioned features, denoted *logMel*, are used directly as input to the classifier model. For the NN architecture, sequentially, a Discrete Cosine Transform (DCT) is applied. The first 14 DCT coefficients were kept, including the 0th order coefficient. Delta (Δ) and acceleration ($\Delta\Delta$) coefficients were also computed, based on a window length of 9. Finally, the mean and standard deviation over time are calculated for each DCT, delta and acceleration feature in the entire audio sample of 15s. The resulting feature vector is of size 84, and referred to as *MFCC*.

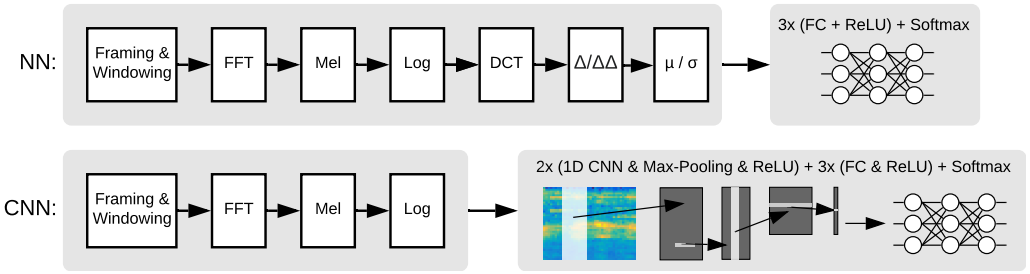


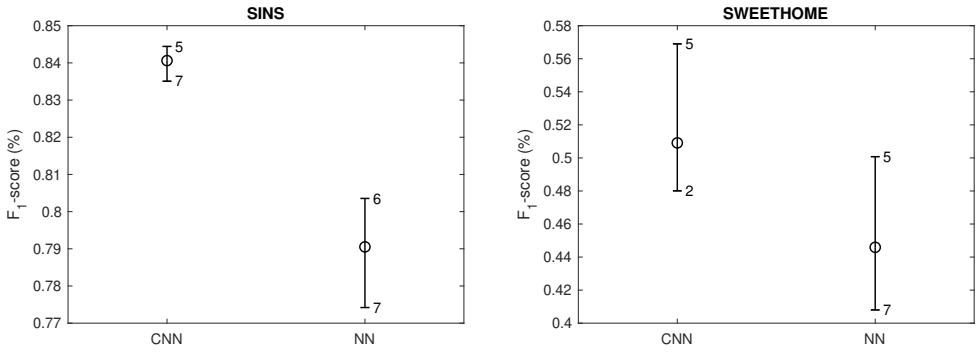
Figure 4: The classifier architectures ranging from low to higher computational complexity: NN and CNN.

Regarding the classification model, *NN* uses the *MFCC* feature vector with size of 84 as input to three Fully Connected (FC) layers, containing 128 neurons, and ReLU activation to allow for non-linear classification. A softmax output layer is used to provide a probabilistic output. The CNN architecture (adopted from the work in [22]) uses the *logMel* features with a size of 26 by 1500. Here, two one-dimensional (1D) convolutional layers are used (i.e. convolution is only performed over the time axis). The first 1D convolutional layer has 64 filters with a kernel size of 26 by 5 and stride 1. Subsampling is then performed by using max pooling of size 1 by 5

and stride 1 by 5. The resulting feature map of size 64 by 299 is then provided to a second 1D convolutional layer that has 128 filters with a kernel size of 64 by 5 and stride 1. Sequentially, this is subsampled by global max pooling to aggregate over the entire time axis. The resulting vector of length 128 is provided to the same network as the NN. All convolutional layers have batch normalisation and ReLU activation at their output. For regularisation, a 20% dropout is used between all convolutional and FC layers. The networks are trained using the Adam optimiser with a learning rate of 0.001 and a batch size of 256 samples. On each epoch, the training dataset is randomly subsampled such that the number of examples for each class match the size of the smallest class. The performance of the model is evaluated every 10 epochs, of 500 in total, on a validation subset. The model with the highest score is used as the final model. As a metric, the macro-averaged F_1-score is used, which is the mean of the class-wise F_1-scores. A single model is trained on data from all sensor nodes. Everything is performed in a cross-validation (CV) fashion where the data is divided into 4 folds without splitting the groups (see Table 1) and keeping a similar class distribution. In each CV iteration two folds are used for training, one for validation and one for testing.

5.3. Results and discussion

Figure 5 shows the F_1-score statistics of the entire set of sensor nodes for both architectures (*CNN* and *NN*) on both datasets. Boundaries for each architecture indicate the best and the worst performing sensor node along with the average in-between (indicated by a circle). The indices next to the maximum and minimum refer to the sensor node index which can be located on Figure 3. On the *SINS* dataset, F_1-scores for the separate sensor nodes range between 83.5%-84.4% for *CNN* and 77.4%-80.35% for *NN*. The variance on the performance for a particular sensor node is lower than on the SWEET-HOME dataset. There the F_1-scores range between 48.0%-56.9% for *CNN* and 40.8%-50.1% for *NN*. This can be explained by the fact that the SWEET-HOME dataset has sensor nodes distributed over multiple rooms. As expected, the best and worst performing sensor nodes, for both datasets, are the sensor nodes closest to all sound sources, while sensor nodes which are further away provide the lowest performance.



(a) (b)

Figure 5: F_1-scores for best, least and average performing sensor node of the CNN and NN architectures on the SINS (a) and SWEETHOME (b) dataset.

Class	CNN	NN
Absence	2 / 1	2 / 3
Calling	6 / 1	6 / 1
Cooking	**4** / 7	**4** / 7
Dishwashing	3 / 4	6 / 4
Eating	**4** / 6	**4** / 1
Other	3 / 4	5 / 1
Vacuumcleaner	6 / 2	3 / 7
Visit	3 / 2	5 / 2
Watching TV	**3** / 5	**3** / 1
Working	**1** / 4	**1** / 4

(a)

Class	CNN	NN
Cleaning	1 / 7	1 / 7
Communication	7 / 1	7 / 2
Dressing	4 / 3	4 / 6
Hygiene	**3** / 7	**3** / 6
Meal	5 / 7	5 / 7
Other	4 / 7	4 / 7
Relaxation	7 / 3	7 / 1
Sleeping	4 / 6	5 / 3

(b)

Table 2: Best (at the left of /) and worst (at the right of /) performing sensor node for each activity in dataset SINS (a) and SWEET-HOME (b).

Table 2 shows the best and worst performing sensor node for each class separately. In both datasets, some activities can be considered to be originated from a fixed area (e.g. watching tv and meal) while others are more uniformly distributed (e.g. absence and vacuum cleaner) over the monitored area. The activities originating from a dense region are expected to be classified better by a nearby sensor node (and vice versa). In the table sensor nodes are indicated in bold if they are the best performing sensor node for that class and are the nearest. Note that, in both datasets, no music or any noise sources were present (e.g. other people) besides sensor and environmental noise. The proximity of the sensor to the audio source correlates well with its classification performance.

6. Decision-level fusion

6.1. Methods

In the previous section, the performance in case of a single sensor node was explored. In this section the performance when fusing information sent by N_n different sensor nodes is analysed. As was motivated in the problem description in Section 3, each sensor node performs local classification, i.e. discriminating the given input signal in one out of N_c pre-defined classes. As a consequence only decision-level fusion strategies are considered which depend on the information provided by the individual sensor nodes, which can be categorised in the following *information levels* [24]:

- *Soft level:* each sensor node provides a N_c-dimensional vector $\mathbf{d}_n = [d_{n,1}, ..., d_{n,N_c}]^\mathsf{T}$ which represents the probability for all classes $c \in \{1, ..., N_c\}$ and a particular node $n \in \{1, ..., N_n\}$,
- *Ranked level:* each sensor node provides a N_r-dimensional vector $\mathbf{r}_n = [r_{n,1}, ..., r_{n,N_r}]^\mathsf{T}$ which consists of labels ($\in \{1, ..., N_c\}$) ranked (high to low) based on probability $\mathbf{d}_n$. In this work this is extended by limiting the amount of labels to be send through to $N_r \leq N_c$. Hereby, it is assumed that the ranking of less probable classes is less relevant for fusion. In case $N_r = N_c$ all labels are send through, similar to the work in [29].

- *Label level:* each sensor node provides a single class label (i.e. the most probable class). This is a special case of the *ranked level* where $N_r = 1$.

To transmit these sources of information, the data needs to be formatted in an efficient manner. Regarding the *soft level* two formats will be compared: a) concatenation of probabilities in unsigned fixed-point 16 bit format of $\mathbf{d}_n$ and b) vector quantised version $Q(\mathbf{d}_n)$ of all probability vectors $\mathbf{d}_n$. While the former quantisation is linear, the latter takes into the account the distribution probabilities for a certain ground truth class. After the model training phase, $N_{vq,c}$ prototype vectors are estimated for a particular ground truth class c using K-means clustering [55] on the training set. To make sure at least one prototype is assigned to a class, vector quantisation is done for each class separately. For transmission, only an index $i_v \in \{0, ..., N_{vq}\}$, with $N_{vq} = \sum_{c=0}^{N_c} N_{vq,c}$, will be transmitted to the fusion center where it will be decoded. For this purpose at least $\log_2(N_{vq}!)$ bits are required to encode this information. Similarly, to send through the *ranked level* information, for a particular rank size N_r the minimal amount of bits is at least $\log_2(\frac{N_c!}{(N_c-N_r)!})$ where an index $i_r \in \{0, ..., N_r\}$ will be transmitted.

In this paper, fusion methods are explored that require no training. Common methods include the product-, sum-, maximum- and minimum-rule in case of *soft level* and the majority vote or borda count in case of *ranked- and label level* [29]. For each information level, all these combination rules were evaluated but no notable differences in terms of classification accuracy were observed. These experiments were not added to the paper. In case of *soft level*, the *mean rule* will be used in the experiments throughout the rest of the paper. The final predicted class in case of the mean rule is formally defined as:

$$\hat{c} = \arg\max_{c \in C} \frac{1}{N_n} \sum_{n=1}^{N_n} d_{n,c} \quad . \tag{2}$$

In case of *ranked level*, the borda count will be used. For each node n, a class is given a particular weight based on the ranking. All weights for each node are added to obtain a final vector that gives the support for each class. The final predicted class in case of borda count is formally defined as:

$$\hat{c} = \arg\max_{c \in C} \sum_{n=1}^{N_n} \sum_{r=1}^{N_r} \omega_r \mathbb{1}\{c = r_{n,r}\} \quad , \tag{3}$$

where $\boldsymbol{\omega} = [\omega_1, ..., \omega_{N_r}]$ denotes the weight vector for a particular ranking and $\mathbb{1}$ the identity which is one in case the argument is satisfied and zero if not. In case of traditional borda count the weight is equal to the amount of classes ranked lower (i.e. $\boldsymbol{\omega}_n = [N_r - 1, N_r - 2, ..., 0]$) [29].

6.2. Results and discussion

In this section the results for different ways to fuse the information received from each sensor node are compared using the same classifier architectures, introduced in Section 5.2, on both datasets. The methods described in previous paragraph using different parameter settings are compared based on classification performance and the amount of communicated bits per classification output. Next to classification performance also a measure related to the autonomy of the WASN nodes is included in the assessment. As was discussed in Section 3, the amount of communicated bits per classification output is used for the latter. However, to preserve a link to a practical solution, estimates of the energy consumption will be given using the energy model

that was introduced in Section 3. Before comparing various methods and settings, first the performance of soft-level fusion at a precision of 16 bit using the *mean-rule*, denoted as *Posterior16*, is shown in function of different node set sizes in Figure 6.

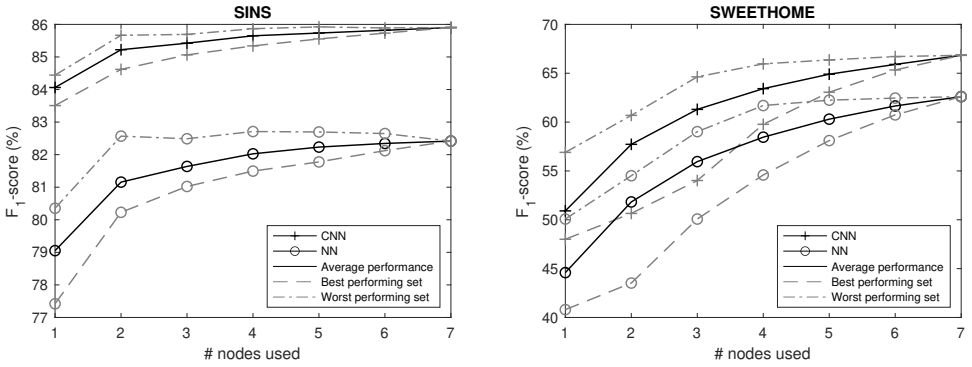


Figure 6: F_1-score with respect to the used sensor node set size in case of mean probability fusion for the CNN and NN architecture on the SINS (a) and SWEET-HOME (b) dataset

In total there are seven sensor nodes, of this set all subsets of size one till seven are evaluated and performance metrics are averaged per set size. This is considered to be an upper bound for fixed decision-level fusion on both the performance and communicated bits as the posterior probabilities of each sensor node are send through in a precision of 16 bit. In case of the *SINS* dataset the average gain in performance from set size of one to seven is, for both NN and CNN, approximately 2%. The difference in performance of the worst and best performing set is relative low (<3%), especially in case of CNN (<1%). As expected, this difference gets smaller with a larger set size. In case of the *SWEET-HOME* dataset the gain in average performance is considerably larger (±15%). Similarly, the difference of the best and worst performing set over all sets is higher (±8%). This is expected as the microphones are deployed over multiple rooms while in case of *SINS* all microphones are in the same room in a relatively small area (± $25m^2$).

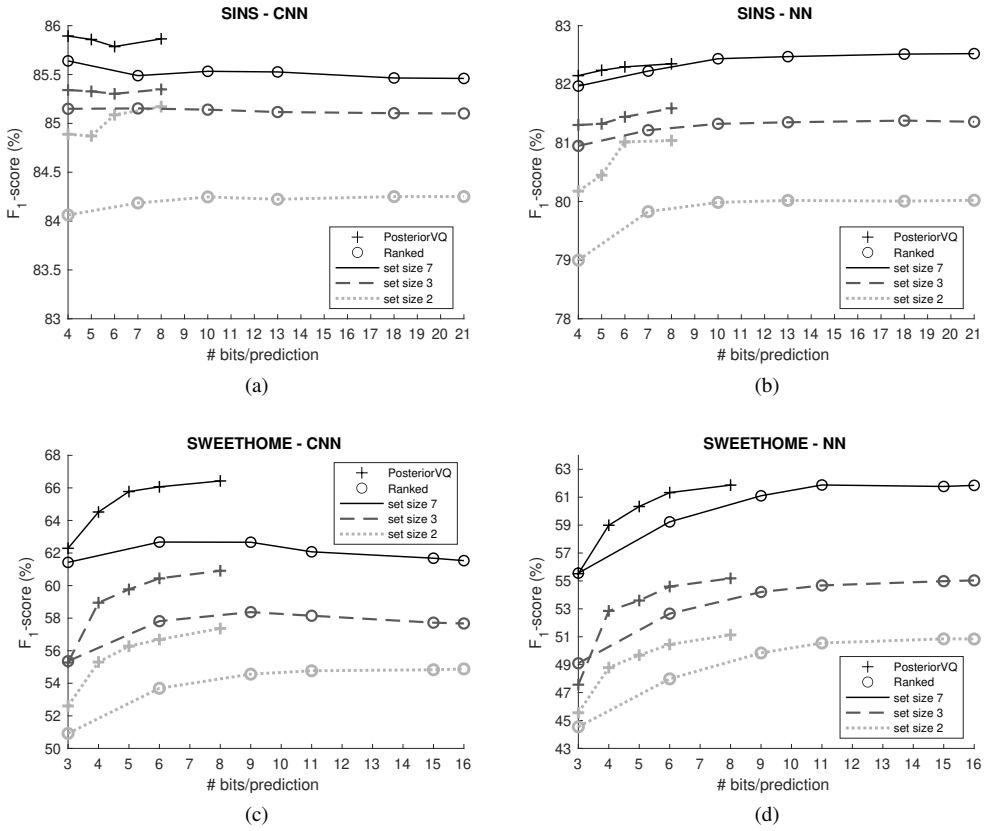


Figure 7: F_1-scores versus amount of communicated bits per prediction for two fusion methods and three set sizes. Each subfigure shows to results for a different dataset and architecture. (a) SINS dataset and CNN architecture, (b) SINS and NN, (c) SWEET-HOME and CNN and (d) SWEET-HOME and NN.

The estimated energy consumption for communicating 16 bit posterior probabilities is 167 μJ per classification output which could be reduced by choosing a different representation and fusion method. Figure 7 shows the F_1-score of two decision-level fusion methods and a range of settings with respect to the amount of communicated bits per classification output. The evaluated methods are *soft level* fusion, with vector quantisation (*PosteriorVQ*) as data representation, and *ranked level* fusion (*Ranked*). Each line in the graph is a function of a parameter that controls the amount of bits. In case of *PosteriorVQ*, this is a function of $N_{vq} \in 2^b$ with $b \in [4, 5, 6, 8]$ for the *SINS* dataset and $b \in [3, 4, 5, 6, 8]$ for the *SWEETHOME* dataset. The difference between the two datasets is due to the amount of classes being 10 and 8 respectively as each class should be represented by at least one prototype vector . Regarding *Ranked*, $N_r \in [1, 2, 3, 4, 6, 8]$. Note that each possible combination for *Ranked* is represented by a single integer and that the amount of bits to represent all these combinations is $\log_2(\frac{N_c!}{(N_c-N_r)!})$, e.g. for $N_r = 8$ at least 16 bits are needed for communicating all different combinations. To improve visibility, the performance is shown for a subset of the possible set sizes (2, 3 and 7 nodes are shown) but this has no effect on the conclusions. For both datasets and architectures it is clear that *PosteriorVQ* outperforms

Ranked. This is particularly the case for smaller set sizes where Ranked suffers from randomness if multiple classes have equal support. In case of the *SINS* dataset the amount of bits can be reduced down to four without losing performance as long as the set size is larger than two. This is because the posterior probabilities are highly skewed towards a single class, i.e. all values are close to zero and close to one for a single class. This is not the case in the *SWEET-HOME* dataset as a loss of performance is noticeable for all set sizes (>4%). In a practical case, the gain in terms of energy consumption is limited due to the used communication standard (i.e. IEEE 802.15.4 [54]). As communication frames are limited to an integer multiple of a single byte, going lower than one byte has no advantage. Additionally, the overhead of a communication frame is typically in the order tens of bytes. When using the energy consumption model introduced in Section 3 the consumed energy for all methods is 126-130 μJ in case of 4-21 bits. With respect to *Posterior16* (160 bit, 167 μJ) there is a decrease in energy consumption of 25%. However, this decrease is negligible compared to the cost at other layers as shown in Figure 2 in Section 3, i.e. the cost for communication ranges in 0.1-1 mJ, while local processing ranges between 10-100 mJ. Note that reducing the amount of bits per classification output could be useful when multiple classification outputs are concatenated in a single communication frame, at the expense of extra latency. To conclude, the benefit of reducing the amount of communicated bits could be useful but strongly depends on the used communication standard and the relative energy cost with respect to other layers.

7. Dynamic sensor activation

7.1. Overview

Dynamic sensor activation could be of interest to reduce the duty cycle of several layers in a sensor node which have a direct impact on the energy costs as indicated in Section 3. In this section three algorithms are proposed compared: 1) locally-controlled, 2) centrally-controlled and 3) hybrid. As the name indicates, locally-controlled refers to an algorithm where each sensor node determines for itself whether or not to communicate, while centrally-controlled gathers all information and decides which are the sensor node(s) that will be active in the next time step. The hybrid form is a combination of both. All dynamic sensor activation strategies are primarily based on two concepts: reliability estimation and best set selection. First, these two concepts are introduced and afterwards the strategies are more elaborately explained. For the remainder of the paper the focus is on vector quantised *soft level* fusion motivated by the previous section.

7.2. Reliability estimation and best set selection

Reliability estimation involves determining if a particular classification output is reliable or not, i.e. a binary classification $\in [0, 1]$, while best set selection is choosing the *best* set of size $N_{n,act}$ out of a set Φ of available sensor nodes that will be made active in the next time step. To enable reliability estimation, Confidence Measures (CM) are introduced. In order to indicate to the user that the system is unsure, CM have been researched in various fields such as automatic speech recognition, image classification and audio detection [32, 56, 57]. When a posterior probability for each class is available an obvious choice is to use this information as CM. Typical approaches are to use the maximum, entropy or N-best likelihood ratio of the posterior probability for each class [56]. Multiple CM methods (e.g. maximum posterior probability [56], N-best likelihood difference [58], relative signal-to-noise ratio [32]) were compared that did not require any training. The maximum element of the posterior probability vector $\mathbf{d}_n$ was selected as it's

easy to interpret and the difference in performance, for the purpose of dynamic sensor activation, between other methods was not significant.

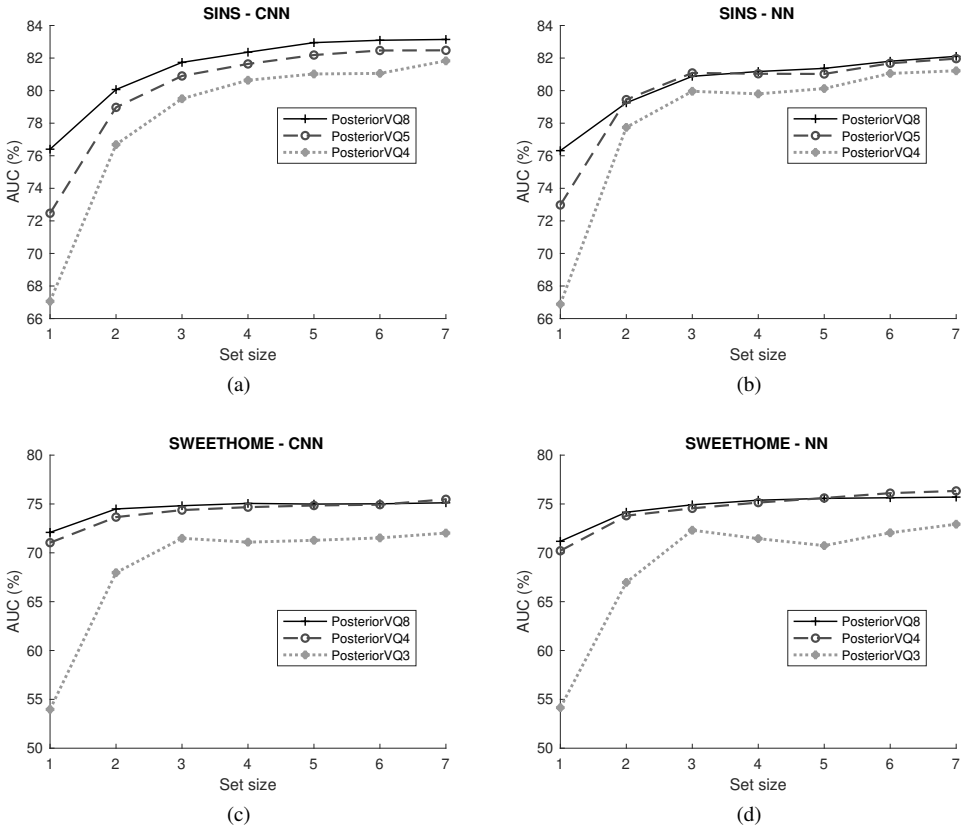


Figure 8: Area Under a Curve of reliability estimation based on maximum posterior probability for various quantisations.

In this work reliability estimation simply involves thresholding a CM, which makes Area Under a Curve (AUC) a natural metric for evaluation. The AUC refers to the total area under the ROC curve, which is set up by plotting the true positive rate in function of the false positive rate. AUC was chosen as a metric at it is threshold independent. Figure 8 shows the AUC for reliability estimation based on the maximum posterior probability versus the set size. The AUC was evaluated for different set sizes, where all possible sensor nodes combinations of that particular set size were averaged. As the experiments contain multiple classes, the average of N_c AUC scores is computed by placing a single class versus all the remaining ones. Each line refers to a different amount of bits used for representing the prototype vectors of *posteriorVQ* including 4, 5 and 8 bits for the *SINS* dataset and 3, 4 and 8 for the *SWEET-HOME* dataset. Lowering the precision has a detrimental effect on the AUC, although this is only noticeable at the lowest amount of bits. Increasing the set size has a positive effect on the AUC, but does not entirely compensate the gap in AUC for the lowest amount of bits.

Regarding best set selection, Figure 9 shows the performance of the set selected with the high-

est reliability (according to the maximum posterior probability rule) along with the performance of selecting a random set for each classification output. Similarly as for reliability estimation, lowering the number of prototype vectors N_{vq} has a detrimental effect and increasing the set size improves the performance. Regarding the *SINS* dataset, the best set has a similar performance, or even better, with respect to using all sensor nodes. In case of the *SWEET-HOME* dataset the decrease in performance from 2^8 to 2^3 prototype vectors is roughly 5%.

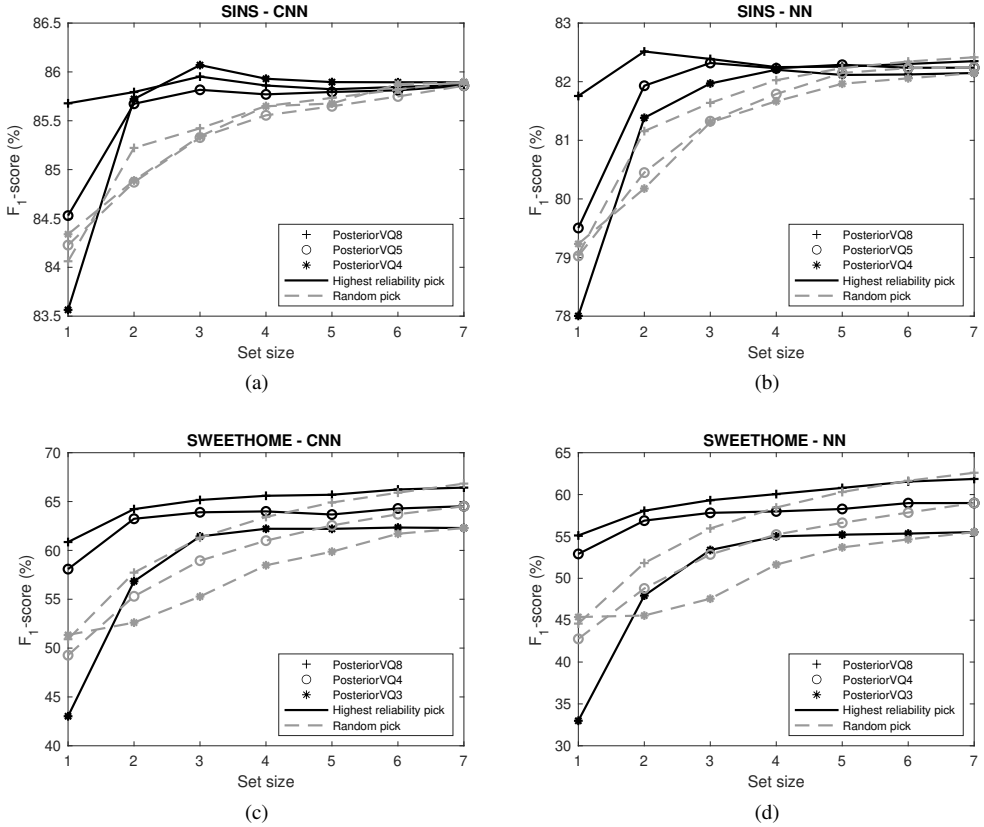


Figure 9: F_1-score with respect to the used sensor node set size for best set selection based on maximum posterior probability and random selection for various quantizations.

7.3. Proposed strategies

The previous paragraph introduced the concept of reliability estimation and best set selection, key elements used in for the dynamic sensor activation strategies introduced here. In this section three strategies are compared that differ on where the decision is made to (de-)activate a sensor, i.e. duty cycling some layers. The sensing layer is considered to be always-on, as the sensor node needs to respond quickly when an activation is desired as classification outputs are based on 15 s. However, the communication layer and/or processing layer could be duty-cycled if needed. Which layers can be duty-cycled depends on the dynamic sensor activation strategy. As mentioned earlier, the strategies primarily differ on where the decision is made to (de-)activate a

layer, i.e. centrally-controlled (D1), locally-controlled (D2) or a combination of both (D3). As these strategies overlap from an algorithmic point of view, a single algorithm for each sensor node and the fusion center is introduced. The algorithm on each sensor node takes care of classification and sends a vector quantised posterior probability in case the current classification output is reliable. Note that the entire sensor node can be (de-)activated by the fusion center. In case of the fusion center, the algorithm takes care of fusing all classification outputs received from a dynamic set of sensor nodes. Additionally, the fusion center can decide to force a sensor node to be active, i.e. classify and transmit. Both algorithms are applicable for the three strategies given a particular parameter setting. Algorithm 1 and Algorithm 2 shows the pseudo-code of the sensor node and fusion center respectively. Regarding Algorithm 1, key parameters include $control \in \{True, False\}$ to allow (de-)activation by the fusion center based on $\phi_n \in \{True, False\}$ and a threshold $\rho_l \in [0, 1]$ on the reliability to control if a vector quantised classification output $Q(\mathbf{d}_n)$ is sent to the fusion center if ϕ_n is False.

Algorithm 1 Sensor node n

Parameters: ρ_l, $control$
Initialise: $\phi_n \leftarrow True$

for every time step **do**
  if $control$ **then** receive ϕ_n
  if ϕ_n **then**
    estimate $\mathbf{d}_n$
    if reliability($\mathbf{d}_n$) $\geq \rho_l$ or ϕ_n **then** transmit $Q(\mathbf{d}_n)$

Regarding Algorithm 2, in case $control$ is True, $N_{n,act} \in \{1, ..., N_n\}$ is the desired set size and $\rho_c \in [0, 1]$ is a threshold on the reliability to either select a new active set Φ or keep the current one.

Algorithm 2 Fusion center

Parameters: $N_{n,act}$, ρ_c, *control*
Initialise: $\Phi \leftarrow \{1, ..., N_n\}$ if *control* else $\emptyset$

for every time step **do**
  for $n \in \{1, ..., N_n\}$ **do**
    if *control* **then**
      $\phi_n \leftarrow$ True if $n \in \Phi$ else False
      transmit ϕ_n to node n
    if received $\mathbf{d}_n$ **then** $\Phi \leftarrow \Phi \cup \{n\}$
  if $|\Phi| > 0$ **then**
    $\overline{\mathbf{d}} \leftarrow \frac{1}{|\Phi|} \sum_{n \in \Phi} \mathbf{d}_n$
    $\hat{c} \leftarrow \arg\max_{c \in C} \overline{d}_c$
  else
    $\hat{c} \leftarrow$ class with highest prior probability
  if *control* **then**
    if reliability($\overline{\mathbf{d}}$) $\geq \rho_c$ and $|\Phi| > N_{n,act}$ **then**
      $\Phi \leftarrow$ BestSetSelection($N_{n,act}$, Φ, $\mathbf{d}$, $\hat{c}$)
    else if reliability($\mathbf{d}_n$) $< \rho_c$ **then** $\Phi \leftarrow \{1, ..., N_n\}$
  else $\Phi \leftarrow \emptyset$

The first strategy for dynamic sensor activation (D1) operates in a locally-controlled manner, and can be obtained by setting *control* to True, ρ_l to 0 and ρ_c is left as an hyperparameter. In that case, all sensor nodes receive $\phi_n \in \{0, 1\}$ from the fusion center, which controls whether to process the data and transmit a classification output. The fusion center fuses the information sent by the set of active sensor nodes $\Phi \in \{1, ..., N_n\}$ to obtain a final classification output. When the reliability is above or equal to threshold ρ_c it will keep the current set active, if not, all sensor nodes are activated to estimate and transmit their vector quantised posterior probability. If the classification outputs of all sensor nodes are reliable, a new set Φ is selected which is described in pseudo-code of Algorithm 3. All possible sets of sensor node combinations S_k in S with $k \in [1, ..., \binom{|\Phi|}{N_{n,act}}]$ given a set size of $N_{n,act}$ are evaluated on their reliability. The most reliable (i.e. best set selection) set, that has a prediction $\hat{c}'$ that equals the prediction $\hat{c}$ of all available sensor nodes, is used as the active set in the next iteration. The sensing layer is always on as the microphone data needs to be buffered in case the sensor node is needed. Additionally, this algorithm adds an additional energy consumption by the sensor nodes since a downlink wireless communication link for ϕ_n is needed to receive control frames.

Algorithm 3 Select active sensor node set

function BESTSETSELECTION($N_{n,act}$, Φ, $\mathbf{d}$, $\hat{c}$)
  $\mathcal{S} \leftarrow$ set of all possible node sets $\mathcal{S}_k$ that satisfy $|\mathcal{S}_k| = N_{n,act}$ and $S_k \in \Phi$
  $\mathcal{R} \leftarrow \emptyset$
  for $\mathcal{S}_k$ in $\mathcal{S}$ **do**
    $\overline{\mathbf{d}} \leftarrow \frac{1}{|\Phi|} \sum_{n \in \Phi} \mathbf{d}_n$
    $\hat{c}' \leftarrow \arg\max_{c \in C} \overline{d}_c$
    if $\hat{c}' = \hat{c}$ **then** $\mathcal{R} \leftarrow \mathcal{R} \cup \{(k, \text{reliability}(\overline{\mathbf{d}}))\}$
  return $\mathcal{S}_k$ with $(k, r) = \arg\max_{(k,r) \in \mathcal{R}} r$

While *D1* was controlled centrally, *D2* is controlled locally. Each sensor node is responsible for determining whether to send something or not. As *control* is set to False, ρ_c does not matter and ρ_l is left as a hyperparameter. Each sensor node performs classification and determines a reliability at every time step. When the reliability of the classification output is above or equal to ρ_l it transmits the classification output to the fusion center. These outputs are fused to obtain the final predicted class. If no sensor nodes send a classification output it chooses the class with the highest prior probability. The algorithm on the sensor node duty cycles the communication layer and lacks the need for reception of frames coming from the fusion center. The processing and sensing layer are always on as the reliability estimate r depends on those.

A disadvantage of *D2* is that it could be that all sensor nodes are not reliable on a local level and send no classification output, while *D1* has a disadvantage of using a quantised version of the posterior probability for the reliability estimation and best selection. *D3*, by setting *control* to True and ρ_c to 0, uses the same principle as *D1* regarding centrally-controlled set selection in case the current classification output is unreliable (i.e. reliability lower than ρ_l). However, all sensor nodes can decide for themselves to send information or not even if ϕ_n is False. Based on the available sensor nodes the fusion center selects a new active set of set size $N_{n,act}$ which sends information regardless of its reliability.

7.4. Results and discussion

Figure 10 and 11 show the results of the proposed strategies in Section 7.3 in which *posteriorVQ* with a precision of 8, 4 and 3 bits respectively was selected as the fusion strategy. Additionally, the average results for static fusion on different set sizes (indicated by text added next to marker $\diamond$) is shown for comparison. As these results are averaged, static fusion can be considered to be similar to the case of random selection of the sensor nodes and is included as a reference. Each curve related to dynamic sensor activation (D1-D3) is a function of the reliability threshold ρ where the marker refers to three distinct (0, 0.5 and 1) values for ρ. Previously these thresholds were denoted as ρ_l and ρ_c, where depending on the used dynamic strategy one of them is set to zero or left as a hyperparameter. In the remainder ρ will be used for simplicity referring to the parameter that is non-zero. Besides the reliability threshold, *D1* and *D2* are also depending on the desired set size $N_{n,act}$. In the figures this is shown for a set size of 1 and 3 (n1 and n3 respectively).

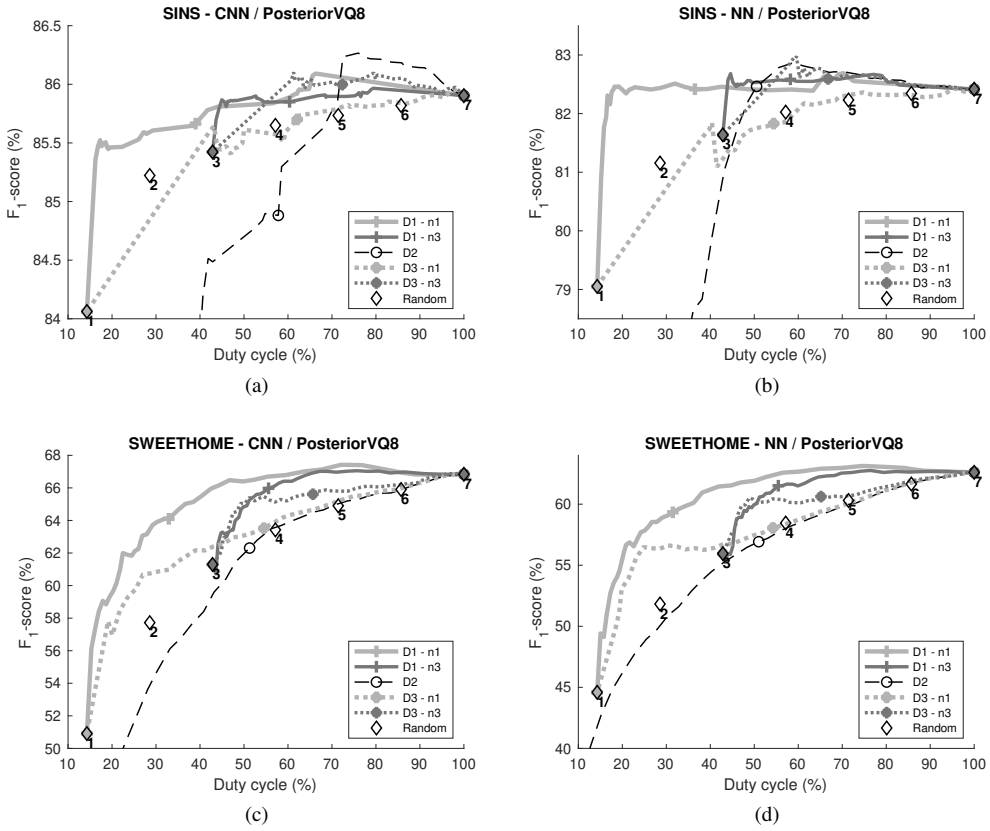


Figure 10: F_1-score with respect to the duty cycle of a sensor node for PosteriorVQ fusion at a precision of 8 bits. Various dynamic sensor activation strategies are shown with respect to the threshold ρ along with random selection of sensor nodes.

Regarding Figure 10, in case of a precision of 8 bits, *D1* is superior to the alternatives for both the *SINS* and *SWEET-HOME* datasets. Roughly speaking, a relative decrease in duty cycle of 50%-80% can be obtained with respect to a static set of sensor nodes without observing a significant different in F_1-score. *D2* suffers from no sensor node being active at lower values of ρ. In case of the *SINS* dataset, using *D1* the duty cycle can be reduced up to 15% without significant loss of performance. Regarding the *SWEET-HOME* dataset the performance is more gradual in function of ρ. The choice of ρ is a trade-off between energy consumption and accuracy. Overall, given these datasets, a choice of $\rho = 0.5$ seems reasonable if similar performance is needed at a lower energy consumption. Regarding the choice of $N_{n,act}$ it is sufficient to limit this to a single active sensor.

Figure 11 shows the results in case of a vector quantised posterior probability at the lowest precision (i.e. 3 and 4 bits for *SWEET-HOME* and *SINS* dataset respectively). Overall, this has a detrimental effect compared to using 8 bits. In case of the *SINS* dataset results are close to randomly selecting sensor nodes. Regarding the *SWEET-HOME* dataset the hybrid approach (*D3*) performs better at lower duty cycles, which is expected as the reliability estimation is not

affected. At higher values of ρ this disadvantage is gone as for *D1* it is more likely to not trust the current classification output and query all sensor nodes to send a classification output, which explains why for higher duty cycles *D1* outperforms *D3*. Unlike the case of a precision of 8 bits, at lower precision the size of the set $N_{n,act}$ has an impact on the F_1-score as adding more sensor nodes increases the performance of best set selection and reliability estimation.

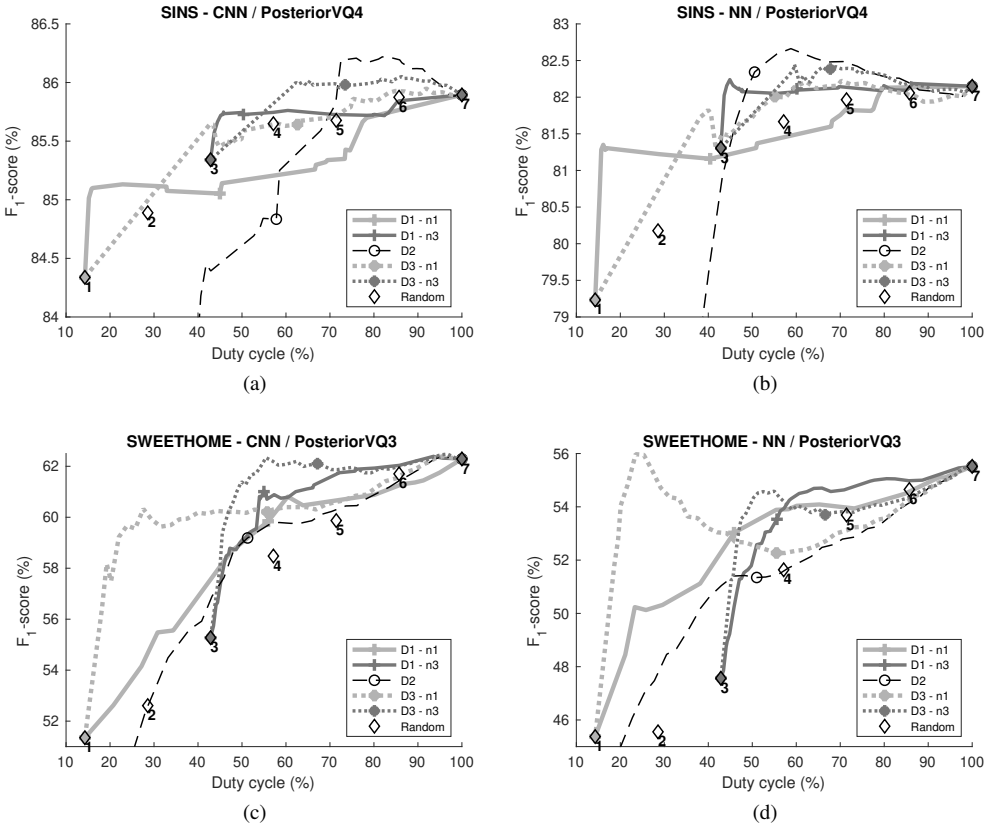


Figure 11: F_1-score with respect to the duty cycle of a sensor node for PosteriorVQ fusion at a precision of 4/3 bits for SINS and SWEET-HOME dataset respectively. Various dynamic sensor activation strategies are shown with respect to the threshold ρ along with random selection of sensor nodes.

Depending on the algorithm, the duty cycle has a different effect on the energy consumption. *D1* has a direct reduction in energy consumption of the processing and wireless transmission, while sensing and wireless reception are always on. *D2* allows for duty cycling on the wireless transmission along with no additional cost for wireless reception. *D3* only introduces duty cycling on the wireless transmission. Depending on the relative energy consumption for each layer a particular scheme may be more favourable than the other. Given a hypothetical hardware architecture and processing, introduced in Section 3 and Paragraph 5.2, *D1* would be the favourable option as the cost for processing is a factor 40-70 more than the other layers, i.e. ±1 mJ for communication versus ±40 mJ and ±70 mJ for architectures *FC* and *CNN* respectively as long as posterior probabilities are represented at sufficient precision. The reduction in energy

consumption is therefore comparable to the decrease in duty cycle, hence resulting in an energy decrease of 50 to 80%.

8. Conclusion

Battery-fed wireless acoustic sensor networks are of interest to many applications, including the use case of classifying daily activities occurring in a home environment. However, commercialisation is limited due to the prohibitively high energy consumption when raw audio data is transmitted. In order to extend the network's lifetime, this paper proposes reducing the amount of communication needed per classification output and dynamically (de-)activating sensor node(s).

First, the energy spend in the sensing, processing and communication layer was compared using an energy consumption model of a hypothetical hardware system that complies with current standards. It was concluded that it is favourable to use decision-level fusion, adopting a topology where processing – including feature extraction and classification – is performed on a (dynamic) set of sensor nodes that output decisions which are fused centrally. Using this topology, a comparison of multiple methods for representing a sensor node's decision indicated that vector quantisation can reduce communication to 8 bit per classification output without significant performance loss. In terms of energy consumption, this resulted in a decrease of up to 3%.

As this is fairly limited, dynamic sensor activation strategies were explored that use a classification output's confidence measure to (de-)activate sensor node(s). These strategies mainly differ by the way sensor (de-)activation is controlled (i.e. centrally, locally or both). It is shown that the algorithm, employing central control, was favoured the most with a reduction in duty cycle down to 20%. As a result this, given the energy consumption model, resulted in an energy decrease of up to 80% as the processing layer dominated the overall energy budget. However, when the amount of bits used for vector quantisation on a sensor node's classification output is limited (1 bit/class), the gains are negligible over random duty cycling. In this case the hybrid approach is preferred but only if the cost for wireless communication dominates the overall energy budget.

Our results suggest that adjusting the likelihood of (de-)activation based on the sensor node's battery depletion rate could bring significant savings; hence, the exploration of such schemes correspond to a promising avenue for future research.

9. CRediT authorship contribution statement

Gert Dekkers: Conceptualisation, Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Writing - original draft, Visualisation, Writing - review & editing. **Fernando Rosas:** Conceptualisation, Methodology, Writing - review & editing. **Toon van Waterschoot:** Conceptualisation, Writing - review & editing, Supervision. **Bart Vanrumste:** Conceptualisation, Writing - review & editing, Supervision. **Peter Karsmakers:** Conceptualisation, Resources, Writing - review & editing, Supervision, Project administration, Funding acquisition.

10. Acknowledgment

This research received funding from the Flemish Government under the "Onderzoeksprogramma Artificiële Intelligentie (AI) Vlaanderen" programme.

References

[1] S. Borkar, A. A. Chien, The future of microprocessors, Commun. ACM 54 (5) (2011) 67–77. doi:10.1145/1941487.1941507.

[2] A. Mainwaring, D. Culler, J. Polastre, R. Szewczyk, J. Anderson, Wireless sensor networks for habitat monitoring, in: Proceedings of the 1st ACM international workshop on Wireless sensor networks and applications, ACM, 2002, pp. 88–97.

[3] I. Butun, S. D. Morgera, R. Sankar, A survey of intrusion detection systems in wireless sensor networks, IEEE Communications Surveys Tutorials 16 (1) (2014) 266–282.

[4] F. Erden, S. Velipasalar, A. Z. Alkar, A. E. Cetin, Sensors in assisted living: A survey of signal and image processing methods, IEEE Signal Processing Magazine 33 (2) (2016) 36–44. doi:10.1109/MSP.2015.2489978.

[5] P. Rawat, K. D. Singh, J. M. Chaouchi, Hakimaand Bonnin, Wireless sensor networks: a survey on recent developments and potential synergies, The Journal of Supercomputing 68 (1) (2014) 1–48. doi:10.1007/s11227-013-1021-9.

[6] I. A. T. Hashem, V. Chang, N. B. Anuar, K. Adewole, I. Yaqoob, A. Gani, E. Ahmed, H. Chiroma, The role of big data in smart city, International Journal of Information Management 36 (5) (2016) 748 – 758.

[7] C. Debes, A. Merentitis, S. Sukhanov, M. Niessen, N. Frangiadakis, A. Bauer, Monitoring activities of daily living in smart homes: Understanding human behavior, IEEE Signal Processing Magazine 33 (2) (2016) 81–94.

[8] A. Bertrand, Applications and trends in wireless acoustic sensor networks: A signal processing perspective, in: 2011 18th IEEE Symposium on Communications and Vehicular Technology in the Benelux (SCVT), 2011, pp. 1–6.

[9] G. Dekkers, S. Lauwereins, B. Thoen, M. W. Adhana, H. Brouckxon, T. van Waterschoot, B. Vanrumste, M. Verhelst, P. Karsmakers, The SINS database for detection of daily activities in a home environment using an acoustic sensor network, in: Proceedings of the Detection and Classification of Acoustic Scenes and Events 2017 Workshop (DCASE2017), Munich, Germany, 2017, pp. 32–36.

[10] M. Vacher, D. Istrate, F. Portet, T. Joubert, T. Chevalier, S. Smidtas, B. Meillon, B. Lecouteux, M. Sehili, P. Chahuara, S. Méniard, The sweet-home project: Audio technology in smart homes to improve well-being and reliance, in: 2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society, 2011, pp. 5291–5294.

[11] D. Stowell, D. Giannoulis, E. Benetos, M. Lagrange, M. D. Plumbley, Detection and classification of acoustic scenes and events, IEEE Transactions on Multimedia 17 (10) (2015) 1733–1746.

[12] A. Mesaros, T. Heittola, E. Benetos, P. Foster, M. Lagrange, T. Virtanen, M. D. Plumbley, Detection and classification of acoustic scenes and events: Outcome of the dcase 2016 challenge, IEEE/ACM Trans. Audio, Speech and Lang. Proc. 26 (2) (2018) 379–393.

[13] A. Mesaros, T. Heittola, A. Diment, B. Elizalde, A. Shah, E. Vincent, B. Raj, T. Virtanen, DCASE 2017 Challenge setup: Tasks, datasets and baseline system, in: Proceedings of the Detection and Classification of Acoustic Scenes and Events 2017 Workshop (DCASE2017), Munich, Germany, 2017.

[14] G. Dekkers, F. Rosas, S. Lauwereins, S. Rajendran, S. Pollin, B. Vanrumste, T. van Waterschoot, M. Verhelst, P. Karsmakers, A multi-layered energy consumption model for smart wireless acoustic sensor networks, Tech. rep., KU Leuven (December 2018). arXiv:1812.06672.

[15] B. Thoen, G. Ottoy, F. Rosas, S. Lauwereins, S. Rajendran, L. De Strycker, S. Pollin, M. Verhelst, Saving energy in wsns for acoustic surveillance applications while maintaining qos, in: 2017 IEEE Sensors Applications Symposium (SAS), 2017, pp. 1–6.

[16] D. Barchiesi, D. Giannoulis, D. Stowell, M. D. Plumbley, Acoustic scene classification: Classifying environments from the sounds they produce, IEEE Signal Processing Magazine 32 (3) (2015) 16–34.

[17] E. Fonseca, M. Plakal, F. Font, D. P. W. Ellis, X. Favory, J. Pons, X. Serra, General-purpose tagging of freesound audio with audioset labels: Task description, dataset, and baseline, in: Proceedings of the Detection and Classification of Acoustic Scenes and Events 2018 Workshop (DCASE2018), Surrey, UK, 2018.

[18] D. Stowell, Y. Stylianou, M. Wood, H. Pamuła, H. Glotin, Automatic acoustic detection of birds through deep learning: the first bird audio detection challenge, in: Proceedings of the Detection and Classification of Acoustic Scenes and Events 2018 Workshop (DCASE2018), Surrey, UK, 2018.

[19] R. Serizel, N. Turpault, H. Eghbal-Zadeh, A. Parag Shah, Large-Scale Weakly Labeled Semi-Supervised Sound Event Detection in Domestic Environments, in: Proceedings of the Detection and Classification of Acoustic Scenes and Events 2018 Workshop (DCASE2018), Surrey, UK, 2018.

[20] M. Vacher, B. Lecouteux, P. Chahuara, F. Portet, B. Meillon, N. Bonnefond, The Sweet-Home speech and multimodal corpus for home automation interaction, in: The 9th edition of the Language Resources and Evaluation Conference (LREC), Reykjavik, Iceland, 2014, pp. 4499–4506.

[21] L. Vuegen, B. Van Den Broeck, P. Karsmakers, H. Van hamme, B. Vanrumste, Energy efficient monitoring of

activities of daily living using wireless acoustic sensor networks in clean and noisy conditions, in: 2015 23rd European Signal Processing Conference (EUSIPCO), 2015, pp. 449–453.
[22] G. Dekkers, L. Vuegen, T. van Waterschoot, B. Vanrumste, P. Karsmakers, DCASE 2018 Challenge - Task 5: Monitoring of domestic activities based on multi-channel acoustics, Tech. rep., KU Leuven (July 2018). arXiv:1807.11246.
[23] B. Khaleghi, A. Khamis, F. O. Karray, S. N. Razavi, Multisensor data fusion: A review of the state-of-the-art, Information Fusion 14 (1) (2013) 28 – 44.
[24] U. Gosa Mangai, S. Samanta, S. Das, P. Roy Chowdhury, A survey of decision fusion and feature fusion strategies for pattern classification, IETE Technical Review 27.
[25] R. C. King, E. Villeneuve, R. J. White, R. S. Sherratt, W. Holderbaum, W. S. Harwin, Application of data fusion techniques and technologies for wearable health monitoring, Medical Engineering and Physics 42 (2017) 1 – 12.
[26] R. Gravina, P. Alinia, H. Ghasemzadeh, G. Fortino, Multi-sensor fusion in body sensor networks: State-of-the-art and research challenges, Information Fusion 35 (2017) 68 – 80.
[27] J. N. Tsitsiklis, M. Athans, On the complexity of decentralized decision making and detection problems, in: The 23rd IEEE Conference on Decision and Control, 1984, pp. 1638–1641.
[28] F. Rosas, K. C. Chen, D. Gündüz, Social learning for resilient data fusion against data falsification attacks, Computational Social Networks 5 (2018).
[29] L. I. Kuncheva, J. Bezdek, R. Duin, Decision templates for multiple classifier fusion: An experimental comparison, Pattern Recognition 34 (2001) 299–314.
[30] J. Kittler, M. Hatef, R. P. W. Duin, J. Matas, On combining classifiers, IEEE Transactions on Pattern Analysis and Machine Intelligence 20 (3) (1998) 226–239.
[31] J. Tsitsiklis, Decentralized detection, in: Advances in Statistical Signal Processing, Vol. 2, 1993, pp. 297–344.
[32] L. Vuegen, B. Van Den Broeck, P. Karsmakers, H. Van hamme, B. Vanrumste, Monitoring activities of daily living using wireless acoustic sensor networks in clean and noisy conditions, in: 2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), 2015, pp. 4966–4969.
[33] R. Grzeszick, A. Plinge, G. A. Fink, Bag-of-features methods for acoustic event detection and classification, IEEE/ACM Transactions on Audio, Speech, and Language Processing 25 (6) (2017) 1242–1252.
[34] I. Martín-Morató, M. Cobos, F. J. Ferri, Analysis of data fusion techniques for multi-microphone audio event detection in adverse environments, in: 2017 IEEE 19th International Workshop on Multimedia Signal Processing (MMSP), 2017, pp. 1–6.
[35] H. Karl, A. Willig, Protocols and architectures for wireless sensor networks, John Wiley & Sons, 2007.
[36] V. Sze, Y. Chen, T. Yang, J. S. Emer, Efficient processing of deep neural networks: A tutorial and survey, Proceedings of the IEEE 105 (12) (2017) 2295–2329. doi:10.1109/JPROC.2017.2761740.
[37] K. M. H. Badami, S. Lauwereins, W. Meert, M. Verhelst, A 90 nm cmos, 6μw power-proportional acoustic sensing frontend for voice activity detection, IEEE Journal of Solid-State Circuits 51 (1) (2016) 291–302.
[38] S. Sigtia, A. M. Stark, S. Krstulović, M. D. Plumbley, Automatic environmental sound recognition: Performance versus computational cost, IEEE/ACM Transactions on Audio, Speech, and Language Processing 24 (11) (2016) 2096–2107.
[39] E. Shih, S.-H. Cho, N. Ickes, R. Min, A. Sinha, A. Wang, A. Chandrakasan, Physical layer driven protocol and algorithm design for energy-efficient wireless sensor networks, in: Proceedings of the 7th annual international conference on Mobile computing and networking, ACM, 2001, pp. 272–287.
[40] D. Feng, C. Jiang, G. Lim, L. J. Cimini, G. Feng, G. Y. Li, A survey of energy-efficient wireless communications, IEEE Communications Surveys Tutorials 15 (1) (2013) 167–178.
[41] Shuguang Cui, A. J. Goldsmith, A. Bahai, Energy-constrained modulation optimization, IEEE Transactions on Wireless Communications 4 (5) (2005) 2349–2360.
[42] F. Rosas, C. Oberli, Energy-efficient MIMO SVD communications, in: 2012 IEEE 23rd International Symposium on Personal, Indoor and Mobile Radio Communications - (PIMRC), 2012, pp. 1588–1593.
[43] D. Ganesan, R. Govindan, S. Shenker, D. Estrin, Highly-resilient, energy-efficient multipath routing in wireless sensor networks, ACM SIGMOBILE Mobile Computing and Communications Review 5 (4) (2001) 11–25.
[44] F. Rosas, R. D. Souza, M. Verhelst, S. Pollin, Energy-efficient MIMO multihop communications using the antenna selection scheme, in: Wireless Communication Systems (ISWCS), 2015 International Symposium on, IEEE, 2015, pp. 686–690.
[45] C. Schurgers, M. B. Srivastava, Energy efficient routing in wireless sensor networks, in: 2001 MILCOM Proceedings Communications for Network-Centric Operations: Creating the Information Force (Cat. No.01CH37277), Vol. 1, 2001, pp. 357–36.
[46] W. Ye, J. Heidemann, D. Estrin, An energy-efficient mac protocol for wireless sensor networks, in: INFOCOM 2002. Twenty-First Annual Joint Conference of the IEEE Computer and Communications Societies. Proceedings. IEEE, Vol. 3, IEEE, 2002, pp. 1567–1576.
[47] H. Yetgin, K. T. K. Cheung, M. El-Hajjar, L. H. Hanzo, A survey of network lifetime maximization techniques in

wireless sensor networks, IEEE Communications Surveys Tutorials 19 (2) (2017) 828–854.
[48] S. Lauwereins, Cross-layer self-adaptivity for ultra-low power responsive IoT devices, Ph.D. thesis, KU Leuven (2018).
[49] G. Dekkers, F. Rosas, WASN EM: a multi-layered Energy Model for Wireless Acoustic Sensor Networks (2018). URL `https://github.com/gertdekkers/WASN_EM/`
[50] F. Rosas, C. Oberli, Modulation and SNR optimization for achieving energy-efficient communications over short-range fading channels, IEEE Transactions on Wireless Communications 11 (12) (2012) 4286–4295.
[51] H. Karl, A. Willig, Protocols and Architectures for Wireless Sensor Networks, John Wiley & Sons, Inc., Hoboken, NJ, USA, 2005.
[52] F. Rosas, G. Brante, R. D. Souza, C. Oberli, Optimizing the code rate for achieving energy-efficient wireless communications, in: Wireless Communications and Networking Conference (WCNC), 2014 IEEE, IEEE, 2014, pp. 775–780.
[53] ARM Limited, Cortex-M4: Technical Reference Manual (March 2010).
[54] Specifications for Local and Metropolitan Area Networks- Specific Requirements Part 15.4 (2006).
[55] S. P. Lloyd, Least squares quantization in PCM, IEEE Transactions on Information Theory 28 (1982) 129–137.
[56] H. Jiang, Confidence measures for speech recognition: A survey, Speech Communication 45 (4) (2005) 455 – 470.
[57] S. Wan, T. Wu, W. H. Wong, C. Lee, Confnet: Predict with confidence, in: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2018, Calgary, AB, Canada, April 15-20, 2018, 2018, pp. 2921–2925.
[58] P. Giannoulis, G. Potamianos, A. Katsamanis, P. Maragos, Multi-microphone fusion for detection of speech and acoustic events in smart spaces, European Signal Processing Conference (2014) 2375–2379.