ENERGY EFFICIENT MONITORING OF ACTIVITIES OF DAILY LIVING USING WIRELESS ACOUSTIC SENSOR NETWORKS IN CLEAN AND NOISY CONDITIONS

Lode Vuegen∗†‡, Bert Van Den Broeck∗†, Peter Karsmakers∗†, Hugo Van hamme‡, Bart Vanrumste∗†§

∗ KU Leuven, Dept. of Electrical Engineering, ESAT-ETC-AdvISe, Kleinhoefstraat 4, B-2440 GEEL, Belgium.
† KU Leuven, Dept. of Electrical Engineering, ESAT-STADIUS, Kasteelpark Arenberg 10, B-3001 LEUVEN, Belgium.
‡ KU Leuven, Dept. of Electrical Engineering, ESAT-PSI, Kasteelpark Arenberg 10, B-3001 LEUVEN, Belgium.
§ KU Leuven, iMinds Future Health Department, Kasteelpark Arenberg 10, B-3001 LEUVEN, Belgium.

E-mail: lode.vuegen@kuleuven.be.

ABSTRACT

This work examines the use of a Wireless Acoustic Sensor Network (WASN) for the classification of clinically relevant activities of daily living (ADL) from elderly people. The aim of this research is to automatically compile a summary report about the performed ADLs which can be easily interpreted by caregivers. In this work the classification performance of the WASN will be evaluated in both clean and noisy conditions. Moreover, the computational complexity of the WASN and solutions to reduce the required computational costs are examined as well. The obtained classification results indicate that the computational cost can be reduced by a factor of 2.43 without a significant loss in accuracy. In addition, the WASN yields a 1.4% to 4.8% increase in classification accuracy in noisy conditions compared to single microphone solutions.

Index Terms— Wireless Acoustic Sensor Networks, health monitoring, activity classification, noise robustness.

1. INTRODUCTION

Due to the baby-boom generation retirement and increasing life expectancy, the ratio of retired to working people is significantly increasing. This aging brings important challenges to our society. One of the main challenges is to assist elderly people to stay as long and safe as possible in their own home environment with minimal personal assistance. This relieves the growing demand for expensive care facilities.

Currently, the golden standard to determine self-reliance of elderly is the Katz index of independence in activities of daily living, often referred to as the Katz ADL [1]. This index measures self-reliance by observing how well following basic tasks are performed: bathing, dressing, toileting, transferring, continence and feeding. The major drawback of this approach is the time and effort required from caregivers to evaluate self-reliance. In addition, this approach to assess the self-reliance is not always objective since it is only a snapshot evaluation.

The aim of this research is to automatically compile a summary report about the performed activities of daily living which can be used by caregivers to assess the health condition more objectively. These reports will be generated based on domestic sounds acquired by a Wireless Acoustic Sensor Network (WASN) installed in the home environment. The use of a WASN for this application has multiple advantages compared to other setups. For instance, the nodes can be small while maintaining large spatial sampling [2]. The nodes can be placed in a room without inconvenient cables. The location of sound sources can be estimated by applying spatial filtering techniques [2]. In addition, the workload can be distributed among nodes, so that relatively inexpensive hardware can be used [2].

The remainder of this paper is organized as follows: Section 2 discusses the developed nodes, the proposed system architecture and the computational complexity of the WASN. Section 3 describes the used classification algorithms together with their computational complexity. The experimental setup and the acquired acoustic dataset are discussed in Section 4. The obtained classification results in both a clean and noisy setup are given in Section 5. Finally, the conclusions and future work are discussed in Section 6.
The required processing power in terms of numerical multipli-
fluence of the following feature parameters is examined w.r.t.
mining the required processing power in each node, the in-
as acoustic features [3].
standard Mel-Frequency Cepstral Coefficients (MFCC) are used
processed in the feature extraction module. In this work stan-
receiving the acoustic data with the highest SNR is further
indicated by the SAD. Only the acoustic data from the node
the average energy in the frames with and without sound as
in the W ASN. The SNR is determined as the ratio between
one of the nodes detects sound in the 30 second block, the av-
assumption that each activity takes at least 30 seconds. When
instead. The block size of 30 seconds is chosen with the
acoustics are useful for activity recognition. Therefore, a
model based SAD since a wide range of
computational cost to train the model parameters will not
the computational complexity of the MFCC algorithm.

2. WIRELESS ACOUSTIC SENSOR NETWORK

2.1. Hardware

Each node in the acoustic sensor network consists of three
linearly spaced MEMS-microphones (SPU0410LR5H) with
an inter-microphone distance of 2.5 cm. The SPU0410LR5H
is a miniature, high-performance, low power microphone well
suited for audio and ultrasound applications. A single-ended
amplifier, with (RF/EMI) protection and a gain-factor of 25.1
dB as advised in the datasheet, was used for pre-amplification
of the sensor signals. All captured acoustic signals were
recorded using a 4 channel 24-bit soundcard sampling at 96
kHz.

2.2. System architecture

The proposed system architecture is shown in Figure 1 and
consists of multiple acoustic nodes as described in Section
2.1. Each node determines in blocks of 30 seconds data
whether or not its input contains acoustic information by
using a Sound Activity Detector (SAD). In this work it is
difficult to use a model based SAD since a wide range of
acoustics are useful for activity recognition. Therefore, a
simple energy based SAD with an adaptive noise floor is used
instead. The block size of 30 seconds is chosen with the
assumption that each activity takes at least 30 seconds. When
one of the nodes detects sound in the 30 second block, the av-
average signal-to-noise (SNR) ratio is estimated for each node
in the WASN. The SNR is determined as the ratio between
the average energy in the frames with and without sound as
indicated by the SAD. Only the acoustic data from the node
receiving the acoustic data with the highest SNR is further
processed in the feature extraction module. In this work stan-
standard Mel-Frequency Cepstral Coefficients (MFCC) are used
as acoustic features [3].

Since computational cost is an important factor for deter-
mining the required processing power in each node, the in-
fluence of the following feature parameters is examined w.r.t.
the required processing power in terms of numerical multipli-
cations: a) sampling frequency, b) number of Mel-filters and
c) number of cepstral coefficients. Table 1 gives an overview
of processing time gain for different parameter settings com-
pared to the baseline setting: sampling frequency of 32 kHz, 20 Mel-filters
and 14 cepstral coefficients.

2.3. Acoustic Classifiers

In this work both Gaussian Mixture Models (GMM) and Sup-
port Vector Machines (SVM) are examined with respect to the
classification of ADLs. The major difference between both
classifiers is that GMMs are based on finding the statistical
properties of the data while SVMs are focusing on finding the
most discriminating properties in the data. The classification
process of both classifiers will be briefly explained in Sec-
tion 3.1 and 3.2 respectively. In addition, an expression for the
computational complexity of both classifiers in recogni-
tion mode will be given as well. It is worth mentioning that the
computational cost to train the model parameters will not
be examined in this work since these are estimated off-line.

3.1. Gaussian Mixture Models (GMM)

Each activity is represented by a set of GMM parameters, de-
noted as $\lambda_1, \lambda_2, \ldots, \lambda_C$, with $c = 1, \ldots, C$ as class labels.
The objective is to find the class model with the Maximum-a-
Posteriori (MAP) probability on a given set of unlabeled test
features $X^{(te)}$. The MAP probability is computed by using

$$
\hat{c} = \arg \max_{c \in C} \frac{p(X^{(te)} | \lambda_c) p(\lambda_c)}{p(X^{(te)})}, 
$$

which can be further reduced into

$$
\hat{c} = \arg \max_{c \in C} p(X^{(te)} | \lambda_c),
$$
since $p(X^{(te)})$ is class independent and due the assumption
of equal class prior probabilities $p(\lambda_c)$ in this work [5].

Equation (2) results in $O(CM N^{(te)} D^2)$ multiplications
with $M$ the number of mixtures in each GMM, $D$ the number
of cepstral coefficients and $N^{(te)}$ the number of test fea-
tures. As already explained in Section 2.2, the WASN per-
forms a classification in blocks of 30 seconds except when none
of the nodes detects sound. This implies that $N^{(te)}$ can
vary between 1 and 3000 depending on the number of frames
detected by the SAD in the corresponding 30 second win-
dow. Therefore, the required number of multiplications for
GMM classification with 10 mixture components ranges be-
tween 19,600 and 58,800,000 for 14 cepstral coefficients and
two between 4,900 and 14,700,000 for 7 cepstral coefficients.

3.2. Support Vector Machines (SVM)

During the classification phase, an unlabeled test feature vec-
tor $x^{(te)}$ is evaluated to the two-class SVM model parameters
by using

$$\hat{c} = \text{sign}(\sum_{i=1}^{N^{(sv)}} \alpha_i K(x^{(te)}, x_i^{(sv)}) + b),$$

(3)

where $x^{(sv)}$ is a support vector, $N^{(sv)}$ is the number of sup-
port vectors, and $K(x^{(te)}, x_i^{(sv)})$ is the Kernel-function which
be seen as a function that describes the similarity between
two feature vectors [6]. Several solutions are presented in the
literature to expand this two-class classification problem into
a multiclass classification problem. Here 1-vs-1 is used as
coding scheme to cope with multiclass problems. This im-
plies that in total $(1/2)C(C-1)$ classifiers are estimated
which distinguish one class from another one. The overall
classification result can then be computed by applying a ma-
ajority voting over the sub-classification results.

In this research SVM uses the mean and variance of each
MFCC dimension as acoustic feature instead of using them
individually like GMM. The mean and variance are computed
from the SAD detected feature frames in each block of 30
seconds data. This implies that the feature dimension is dou-
bled compared to the GMM approach and that $N^{(te)}$ is always
equal to one except when none of the nodes in the WASN de-
tects sound in the corresponding window. Equation (3) results
therefore in $O(C(C-1) N^{(sv)} D)$ multiplications with $N^{(sv)}$
the number of support vectors. The number of support vectors
is estimated from the size of the training dataset and is equal
to 270 in this work. This makes that SVM requires in total
680,400 and 340,200 multiplications for the classification of
features with 14 and 7 cepstral coefficients respectively.

4. EXPERIMENTAL SETUP

4.1. Living environment and recorded dataset

Figure 2 presents the floor map of the home environment used
for the recordings of daily living activities. In total seven dif-
ferent nodes, each marked by a red rectangular box with an
arrow indicating the orientation, were placed in the home en-
vironment at a height of approximately 1.75 m. Four of the
seven were placed in each corner of the combined living room
and kitchen. The remaining three were placed in the bedroom,
bathroom and toilet. This implies that each room of the envi-
ronment was covered during the experiments.

In total 10 different activities were performed multi-
ple times by two test users and recorded by the WASN.
These activities were chosen such that these are related to
the Katz scale of independence and are: "Brushing theeth",
"Dishes", "Dressing", "Eating", "Preparing food", "Set-
ting table", "Showering", "Sleeping", "Toileting" and
"Washing hands".

4.2. Simulation environment

A simulation environment was used to create an artificial
noise dataset to examine the influence of background noise
on the classification performance of the WASN [7]. This sim-
ulation environment estimates the room impulse responses
(RIRs) from a particular noise source location to each mi-
crophone in the WASN on basis of the $T_{60}$ time, the room
dimensions and the microphone positions and orientations.
All these parameters were measured during the installa-
tion of the WASN to parameterize the simulation model. In this
research the publicly available CHiME dataset was used for
this task [8]. This dataset contains clean examples of typical
domestic noise sources such as speech, television and radio. This noise dataset can be filtered by the RIRs to generate an artificial background noise dataset. In total two different noise source positions, each marked by a blue circle in Figure 2, will be examined in this work. The position of the noise sources is chosen such that each noise source is located approximately 35 cm to one of the nodes in the WASN. The latter is done to examine if the WASN yields better classification accuracies in noisy situations compared to single microphone solutions.

5. RESULTS

During the experiments, the maximum-likelihood (ML) parameters of the GMM models are estimated using expectation-maximization (EM) on the training data as proposed in [5]. Previous research indicates that GMMs with 10 mixture components are typical sufficient for the classification of activities of daily living [9]. Therefore, each GMM in this work is modeled with 10 mixture components. The SVM hyperparameters on the other hand are selected by applying a cross-validation in the training dataset. In this research a radial basis function (RBF) was used as kernel. This implies that during each fold a grid search over the trade-off parameter C and the kernel parameter $\gamma$ is performed to find their optimal value. The creation of the training and test set in this work was done by applying a 2-fold cross-validation two times which results in four equally sized training and test sets for the experiments.

5.1. Clean data

The obtained classification results for both GMM and SVM on the clean dataset are given in Tables 2 and 3 respectively. These results indicate that GMM and SVM are equivalent in classifying activities from acoustic sensor data. However, the computational complexity of SVM is most of the time lower compared to GMM due to the large number of SAD detected test features in the 30 second windows. In addition, these results also indicate that a sampling frequency of 16 kHz is sufficient for ADL classification since lowering the sampling frequency to 8 kHz yields a decrease in accuracy. Increasing the sampling frequency from 16 kHz to 32 kHz on the other hand results only in a slight increase in accuracy while the re-

<table>
<thead>
<tr>
<th># Mel-filters</th>
<th># cepstral coefficients</th>
<th>Sampling frequency 8 kHz</th>
<th>16 kHz</th>
<th>32 kHz</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>7</td>
<td>69.6±3.3%</td>
<td>73.3±4.4%</td>
<td>73.6±5.2%</td>
</tr>
<tr>
<td>15</td>
<td>7</td>
<td>70.4±4.2%</td>
<td>73.4±4.8%</td>
<td>74.2±5.3%</td>
</tr>
<tr>
<td>15</td>
<td>14</td>
<td>72.8±4.8%</td>
<td>75.1±4.5%</td>
<td>76.5±4.8%</td>
</tr>
<tr>
<td>20</td>
<td>7</td>
<td>70.2±3.1%</td>
<td>72.8±4.9%</td>
<td>74.2±5.3%</td>
</tr>
<tr>
<td>20</td>
<td>14</td>
<td>72.7±4.4%</td>
<td>75.5±5.1%</td>
<td>73.0±4.7%</td>
</tr>
</tbody>
</table>

Table 2. Clean GMM classification results for the different feature parameter settings.

<table>
<thead>
<tr>
<th># Mel-filters</th>
<th># cepstral coefficients</th>
<th>Sampling frequency 8 kHz</th>
<th>16 kHz</th>
<th>32 kHz</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>7</td>
<td>68.3±5.5%</td>
<td>72.9±1.7%</td>
<td>71.4±2.8%</td>
</tr>
<tr>
<td>15</td>
<td>7</td>
<td>69.3±5.9%</td>
<td>72.8±4.0%</td>
<td>73.5±2.0%</td>
</tr>
<tr>
<td>15</td>
<td>14</td>
<td>72.8±5.1%</td>
<td>78.0±2.8%</td>
<td>76.9±2.8%</td>
</tr>
<tr>
<td>20</td>
<td>7</td>
<td>70.2±7.4%</td>
<td>72.7±0.7%</td>
<td>71.3±2.4%</td>
</tr>
<tr>
<td>20</td>
<td>14</td>
<td>69.3±2.7%</td>
<td>75.3±4.3%</td>
<td>78.2±4.1%</td>
</tr>
</tbody>
</table>

Table 3. Clean SVM classification results for the different feature parameter settings.
quired time for the feature extraction is doubled. Therefore, SVM with a sampling frequency of 16 kHz is preferred over the other setups and will be examined further in noisy conditions.

5.2. Noisy data

Figures 3 and 4 present the results when a background noise source was inserted in the living environment at position A and B respectively. During these experiments, the loudness of the noise source was set at different levels such that the average SNR over all nodes ranges between 30 dB and 0 dB. The latter is done to examine the influence of background noise on the classification performance of the WASN. In addition, the obtained classification scores of each single node in the combined living room and kitchen instead of selecting the one with the highest SNR are given as well. These results are used to examine if the WASN yields better classification accuracies compared to single microphone solutions. It is worth mentioning that during these experiments the same SAD indices are used as in the clean data. The latter is done to eliminate the influence of incorrect sound activity detection on the classification performance of the WASN.

The results obtained indicate that selecting the node with the highest SNR in the combined living room and kitchen results in higher classification accuracies for medium and high SNRs. However, for very low SNRs, i.e. 10 dB or less, selecting the node with the highest SNR no longer yields better classification accuracies. The latter can be explained by the fact that in severe noisy conditions the acoustic information received by the node with the highest SNR is masked with background noise as well. On the other hand, Figures 3 and 4 indicate that the WASN also has also a slightly poorer classification performance in noise free conditions compared to single microphone solutions. These lower WASN accuracies in a clean setup can be explained in all probability due to the presence of different types of sensor noise for each node which affects the classification performance.

6. CONCLUSIONS AND FUTURE WORK

In this paper the performance of the WASN is examined for the purpose of classification of activities of daily living w.r.t. computational complexity and noise robustness. The highest classification score in clean conditions was obtained for the SVM classifier for a sampling frequency of 32 kHz, 20 Mel-filters and 14 cepstral coefficients. This parameter setting results in a classification accuracy of 78.2 ± 4.1%. However, by halving the sampling frequency to 16 kHz and reducing the number of Mel-filters to 15 results only in an accuracy decrease of 0.2% while the required computational complexity is reduced by a factor of 2.43. In addition, the experiments performed under noisy conditions indicate that a WASN yields better classification results compared to single microphone solutions. The average increase in accuracy for high to medium SNRs ranges between 1.4% and 4.8% and between 1.6% and 3.7% when the noise source was set at positions A and B respectively.

Future work will include spatial features for the classification of ADLs. It can be assumed that including spatial information as a feature might improve the classification performance of the WASN. The latter can be explained by the fact that some activities are always performed at a consistent place in the home environment such as toileting or showering. In addition, the WASN will also be validated on a larger and real-life dataset recorded at the home of an elderly person.

REFERENCES