Monitoring Activities of Daily Living Using Wireless Acoustic Sensor Networks in Clean and Noisy Conditions

Lode Vuegen[1,2,3], Bert Van Den Broeck[1,2], Peter Karsmakers[1,2], Hugo Van hamme[3] and Bart Vanrumste[1,2,4]
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) of 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. Results indicate that the classification performance of the WASN is $75.3 \pm 4.3\%$ on clean acoustic data selected from the node receiving with the highest SNR. By incorporating spatial information extracted by the WASN, the classification accuracy further increases to $78.6 \pm 1.4\%$. In addition, the classification performance of the WASN in noisy conditions is in absolute average 8.1% to 9.0% more accurate compared to highest obtained single microphone results.**

I. 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 alone living 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. In addition, this approach is not always objective since it is only a snapshot evaluation.

The aim of this research is to automatically compile a report that summarizes the activities of daily living performed by the elderly 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. These WASNs have advantages over other kinds of 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 II discusses the developed nodes and the proposed system architecture of the WASN. Section III describes the extracted features and the used classification algorithm while the experimental setup and the acquired acoustic dataset are discussed in Section IV. The obtained classification results in both a clean and noisy setup are given in Section V. Finally, the conclusions and future work is discussed in Section VI.

II. WIRELESS ACOUSTIC SENSOR NETWORK

A. Hardware

Each node in the acoustic sensor network consists of three linearly spaced MEMS-microphones (SPU0410LR5H) with an inter-sensor 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 application note, was used for pre-amplification of the sensor signals. All captured acoustic signals were recorded using a 4 channel 24-bit soundcard sampling at 16 kHz.

B. System architecture

The proposed system architecture is shown in Figure 1 and consists of multiple acoustic nodes as described in Section II-A. Each node determines whether or not its input contains acoustic information by using a sound activity detector

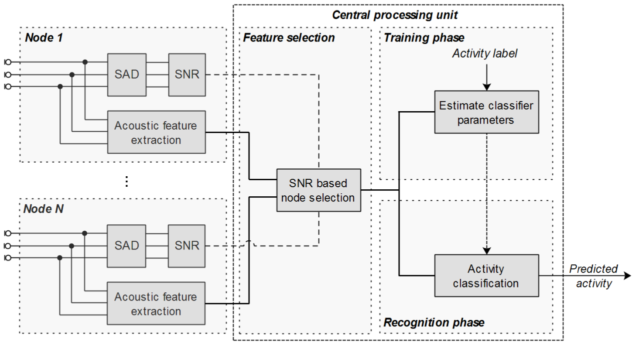

Fig. 1. System architecture of the WASN.

[1] KU Leuven, Department of Electrical Engineering, ESAT-ETC-AdvISe, Kleinhoefstraat 4, B-2440 GEEL, Belgium.

[2] KU Leuven, Department of Electrical Engineering, ESAT-STADIUS, Kasteelpark Arenberg 10, B-3001 LEUVEN, Belgium.

[3] KU Leuven, Department of Electrical Engineering, ESAT-PSI, Kasteelpark Arenberg 10, B-3001 LEUVEN, Belgium.

[4] KU Leuven, iMinds Future Health Department, Kasteelpark Arenberg 10, B-3001 LEUVEN, Belgium.

(SAD) for every block of 30 seconds data. This block size is chosen with the assumption that each activity takes at least 30 seconds. In this work, a simple energy based SAD with an adaptive noise floor is used instead of a model based SAD since a wide range of acoustics are useful for activity recognition. If one of the nodes detects sound, the average signal-to-noise (SNR) ratio of the corresponding block is estimated for each node in the WASN. This by taking the ratio between the average energy of the SAD detected feature frames and the non-SAD detected feature frames. Only the data from the node with the highest SNR will be further processed by the WASN for the classification task.

III. Feature extraction and modeling

A. Feature extraction

Most of the presently available acoustic feature extraction approaches find their origin in speech applications and are often based on the properties of human speech production and perception. A well-known and frequently used feature extraction approach in the domain of speech and speaker recognition applications are the so-called Mel-Frequency Cepstral Coefficients (MFCCs) [3]. Despite the fact that MFCCs are initially developed for speech applications, research indicates that MFCCs are a successful choice for processing non-speech acoustic signals as well [4]. Therefore, MFCCs will be extracted from the acoustic sensor data acquired by the WASN.

In addition, the ID of the selected node, which can be interpreted as a rough sound source localization (SSL), is also used as a feature. The latter is done to examine if position information might boost the classification performance of the WASN because some activities contain some correlated acoustic information. For instance, running water detected in the bathroom is associated more to personal hygiene than to cooking.

B. Acoustic classifier

In this work, a Support Vector Machine (SVM) is used for the purpose of ADL classification. This type of classifier is based on finding a decision function in the feature space which maximizes the margin between two classes. Consider a binary classification task where the goal is to learn a decision function between two activities based on a training set that consists out of N observations $\mathscr{D} = \{(x_i, y_i)\}_{i=1}^{N}$ with measurements $x_i \in \mathbb{R}^D$ and corresponding output class labels $y_i \in \{-1, 1\}$. In a SVM the decision function is modeled by a function $f(x) = w^T\varphi(x) + b$ with $\varphi : \mathbb{R}^D \rightarrow \mathbb{R}^{D_\varphi}$ which is a fixed but unknown mapping. Function $f(x)$ is linear in the parameters $w \in \mathbb{R}^{\varphi_D}$ and $b \in \mathbb{R}$ which are estimated by minimizing the following optimization criterion,

$$\begin{aligned} &\min_{w,b,\xi_i} ||w||_2^2 + C\sum_{i=1}^{N}\xi_i \\ &\quad\text{such that} \\ &\quad y_i(w^T\varphi(x_i)+b) \geq 1-\xi_i, \forall i = 1,...,N \\ &\quad \xi_i \geq 0, \forall i = 1,...,N, \end{aligned} \tag{1}$$

where ξ_i is a slack variable that relaxes the inquality constraint on the latent values $f(x_i)$ and $C \in \mathbb{R}^+\backslash\{0\}$ is the trade-off parameter. The latter balances between maximizing the margin (first term in (1)) and minimizing the number of classification errors that are made on the training set (second term in (1)). The prediction of the label corresponding to a novel example x^* can be determined by:

$$\hat{y}^* = \text{sign}(f(x^*)) \tag{2}$$

Transforming (1) to its equivalent dual formulation and writing the solution in terms of Lagrange multipliers $\alpha_i \geq 0$ gives:

$$\begin{aligned} &\max_{\alpha} \sum_{i=1}^{N}\alpha_i - 1/2\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j K(x_i, x_j), \\ &\quad\text{such that} \\ &\quad \sum_{i=1}^{N}\alpha_i y_i = 0 \\ &\quad 0 \leq \alpha_i \leq C, \forall i = 1,...,N, \end{aligned} \tag{3}$$

where $K(x,x') = \varphi(x)^T\varphi(x')$ is the kernel function. A popular choice for such kernel function is either the linear kernel $(K(x,x') = x^Tx')$ or the Radial Basis Function (RBF) kernel $(K(x,x') = exp(-||x-x'||_2^2/2\sigma^2))$ where hyper-parameter σ defines the bandwidth of this kernel.

Due to the specific type of relaxation applied on the inequalities in (1) some α_i values are driven to zero. In the resulting classifier model the sum should be taken only over the non-zero α_i values instead of all training examples which gives:

$$\hat{y}^* = \text{sign}\left(\sum_{i=1}^{N_{sv}}\alpha_i y_i K(x, x_i) + b\right), \tag{4}$$

where index i runs now over the number of SVs denoted as N_{SV}.

Several solutions are presented in the literature to expand this two-class classification problem into a G multiclass classification problem. Here, 1-vs-1 is used as coding scheme to cope with multiclass problems. This implies that in total $(1\backslash2)G(G-1)$ classifiers are estimated which distinguish one class from another one. The overall classification result can then be computed by applying a majority voting over the sub-classification results.

In this work, the SVM uses the normalized mean and variance of each MFCC dimension as features. The mean and variance are computed from the SAD detected feature frames in each 30 second block. The latter is done to reduce the computational complexity of the SVM significantly since there is linear dependency of complexity on the number of training features [5]. This implies that in recognition mode a classification is done on each 30 second block.

IV. Experimental setup

A. Living environment and recorded dataset

Figure 3 is the floor map of the used home environment for the recording of activities of daily living. In total seven

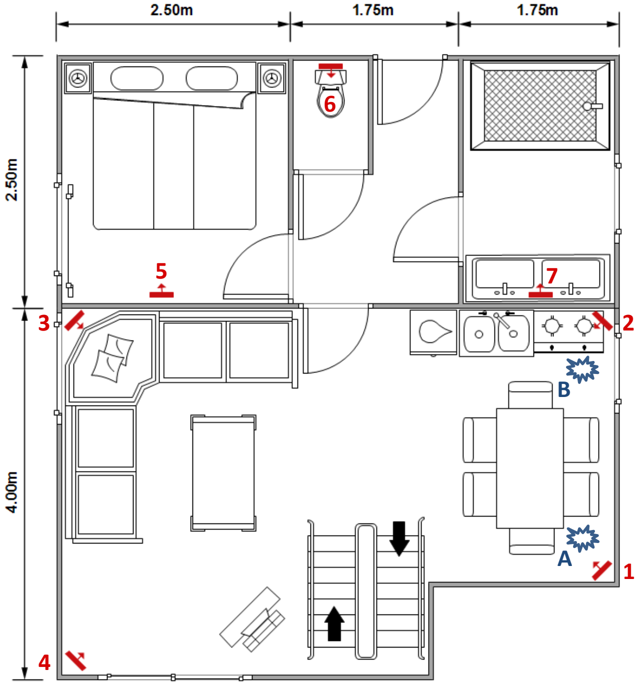

Fig. 2. Living environment with (a) the node positions indicated by numbers 1 to 7 and (b) the artificial noise source positions indicated by A and B.

TABLE I

RECORDED DATASET IN THE LIVING ENVIRONMENT.

Activity	ID	# examples	Duration (min.)
Brushing teeth	1	6	2.0 ± 0.2
Dishes	2	4	7.0 ± 0.6
Dressing	3	24	1.5 ± 0.6
Eating	4	10	4.9 ± 1.2
Preparing food	5	4	8.7 ± 1.7
Setting table	6	6	4.8 ± 1.0
Showering	7	6	2.6 ± 0.6
Sleeping	8	6	3.1 ± 0.4
Toilet	9	12	1.2 ± 0.1
Washing hands	10	6	0.9 ± 0.1

different nodes, each marked by red rectangular box with an arrow indicating the orientation, were placed in the home environment at an height of approximately 1.75 m. Four of the seven were placed in each corner of the open plan living and kitchen area. The remaining three nodes were placed in the bedroom, bathroom and toilet respectively. This implies that each room of the environment was covered during the experiments.

In total 10 different activities were recorded in the living environment. These activities were performed by two test users multiple times and were chosen such that these are related to the Katz scale of independence. An overview of the recorded dataset can be found in Table I.

B. 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 [6]. This simulation environment estimates the room impulse responses (RIRs) from a particular noise source location to each microphone in the WASN on basis of the T60 time, the room dimensions and the microphone positions and orientations. All these parameters were measured during the installation of the WASN to parameterize the simulation model. In this work, the publicly available CHiME dataset was used for this task [7]. This dataset contains clean examples of typical domestic noise sources such as speech, television and radio. This noise dataset can be filtered by the obtained RIRs to generate an artificial background noise dataset. In total two different noise source positions, each marked by a blue circle in Figure 3, 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.

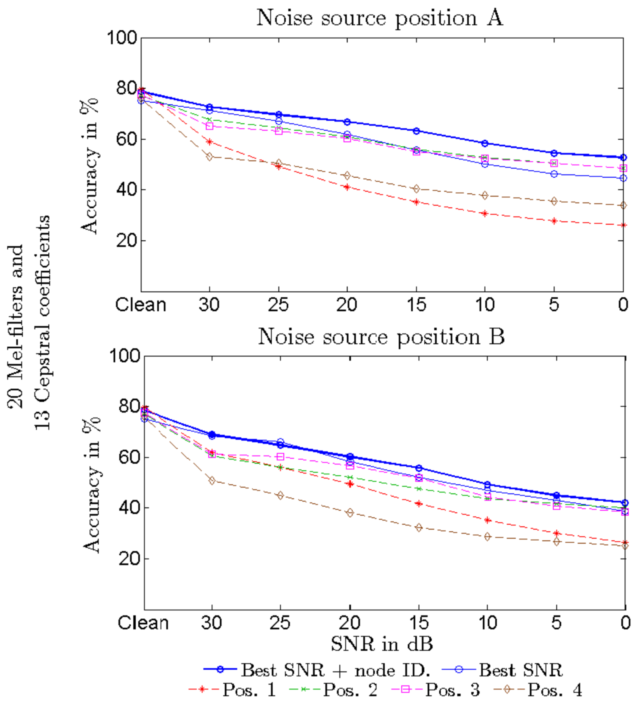

Fig. 3. SVM classification results with noise source at position A and B.

V. RESULTS

During the experiments, the optimal SVM hyperparameters 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 also that during each fold a grid search over the trade-off parameter C and the kernel hyperparameter σ is performed to find their optimal value.

The obtained classification results when the background noise source was set at position A and B are given in Figure 3. During these experiments, the loudness of the background noise source was set at different levels such that the overall SNR over all nodes ranged between 30 dB and 0 dB. To

TABLE II

RESULTS BEST SNR WITHOUT NODE ID: 75.3 ± 4.3%.

		Classified label									
		1	2	3	4	5	6	7	8	9	10
Ground truth	1	**44**	2	-	-	-	-	2	-	-	-
	2	2	**65**	10	22	9	8	-	-	-	-
	3	-	5	**124**	4	1	3	-	1	-	-
	4	1	20	9	**159**	6	5	4	2	-	-
	5	-	30	9	16	**74**	10	1	2	-	-
	6	1	13	12	4	8	**81**	-	3	-	-
	7	2	-	-	-	-	-	**63**	-	-	-
	8	-	-	4	6	1	-	-	**29**	-	-
	9	-	-	-	-	1	-	-	-	**71**	-
	10	-	-	1	-	-	-	1	-	-	**22**

TABLE III

RESULTS BEST SNR WITH NODE ID: 78.6 ± 1.4%.

		Classified label									
		1	2	3	4	5	6	7	8	9	10
Ground truth	1	**44**	-	-	-	-	-	3	-	-	1
	2	2	**72**	-	21	12	9	-	-	-	-
	3	-	-	**136**	-	-	-	-	-	-	-
	4	-	23	3	**166**	6	8	-	-	-	-
	5	-	29	2	13	**83**	10	-	5	-	-
	6	-	16	3	6	13	**81**	-	3	-	-
	7	2	-	-	-	-	-	**63**	-	-	-
	8	-	1	6	4	2	-	-	**27**	-	-
	9	-	-	-	-	-	-	-	-	**72**	-
	10	-	-	1	-	-	-	2	-	-	**21**

examine noise free conditions, the background noise source was switched off completely. In addition, to investigate if the WASN yields higher classification accuracies compared to single microphone solutions, the obtained classification results when only data from one of the four nodes in the open plan living and kitchen area is used instead of selecting the node with the highest SNR, are given as well. It is worth mentioning that during the noisy experiments the same SAD indexes are used as in the clean data to eliminate the influence of incorrect sound activity detection on the classification performance of the WASN.

From the obtained results it can be seen that selecting the node with the highest SNR results in higher classification accuracies for medium and high SNRs compared to single microphone solutions. However, for very low SNRs, i.e. 10 dB or less, SNR-based node selection no longer yields better classification accuracies. This effect 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, Figure 3 also indicates that the WASN has a slightly poorer classification performance in noise free conditions compared to single microphone solutions. These lower WASN accuracies are probably due to the presence of different types of sensor noise which becomes more dominant in noise free conditions and thereby affecting the classification performance of the WASN.

Moreover, these results also indicate that including the ID of the node with the highest SNR as a sound source localization feature results into 3.3% more accurate classification results with 2.9% less variation for the clean setup. By examining the corresponding confusion matrices, which are given in Table II and Table III respectively, it can be seen that this increase in accuracy is due to the better classification performance of the activities: *"Dishes"*, *"Dressing"*, *"Eating"* and *Preparing food*. In addition, the average increase in accuracy for noisy conditions compared to the single microphone results is even larger and varies between 8.1% and 9.0%.

VI. CONCLUSIONS AND FUTURE WORK

In this paper, the performance of the WASN is examined for the purpose of classification of activities of daily living in both clean and noisy setups. The obtained results indicates that a WASN and single microphone solutions are more or less comparable in terms of classification performance in noise free conditions. However, this is not the case any more in noisy conditions where the WASN outperforms single microphone solutions. In addition, including the selected node ID as spatial feature boosts the classification performance of the WASN further.

Future work will include more advanced spatial features for the classification of ADLs. It can be assumed that including direction of arrival information as a feature might further improve the classification performance of the WASN. In addition, the WASN will also be validated on a larger and real-life dataset recorded at the home of an elderly person.

ACKNOWLEDGMENT

This work was performed in the context of following projects: IWT doctoral scholarships (contract 111433 and 121565), Sound INterfacing through the Swarm - SINS (IWT-SBO contract 130006), Algorithms, Architectures and Platforms for Enhanced Living Environments - AAPELE (FP7-COST Action IC1303) and Profound (EC-ICT-PSP contract 325087).

REFERENCES

[1] S. Katz, "Assessing self-maintenance: activities of daily living, mobility, and instrumental activities of daily living," *Journal of the American Geriatrics Society*, vol. 31, no. 12, pp. 721–727, Decmber 1983.

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

[3] D.A. Reynolds and R.C. Rose, "Robust text-independent speaker identification using gaussian mixture speaker models," *Speech and Audio Processing, IEEE Transactions on*, vol. 3, no. 1, pp. 72–83, Jan 1995.

[4] D. Giannoulis, E. Benetos, D. Stowell, M. Rossignol, M. Lagrange, and M. D. Plumbley, "Detection and classification of acoustic scenes and events," in *WASPAA*, 2013, pp. 1–4.

[5] B. E. Boser, I. M. Guyon, and V. N. Vapnik, "A training algorithm for optimal margin classifiers," in *Proceedings of the 5th Annual ACM Workshop on Computational Learning Theory*. 1992, pp. 144–152, ACM Press.

[6] E. A. P. Habets, "Room impulse response generator," Tech. Rep., Technische Universiteit Eindhoven, Eindhoven, 2010.

[7] J. Barker, E. Vincent, N. Ma, H. Christensen, and P. Green, "The PASCAL CHiME Speech Separation and Recognition Challenge," *Computer Speech and Language*, vol. 27, no. 3, pp. 621–633, Feb. 2013.