Detection and Classification of Acoustic Events for In-Home Care

Jens Schroeder, Stefan Wabnik, Peter W.J. van Hengel, and Stefan Goetze

Fraunhofer IDMT, Project group Hearing Speech and Audio (HSA), Oldenburg, Germany
{Jens.Schroeder,Stefan.Wabnik,Peter.vanHengel,
S.Goetze}@idmt.fraunhofer.de

Abstract. Due to the demographic change, the number of older people will grow steadily within the next decades [1]. Technical systems, methods and algorithms developed in the context of Ambient Assisted Living (AAL) aim to enable these people to live self-determined and safe in their own homes as long as possible. To ensure this, these homes may have to be equipped with e.g., a health monitoring system, special sensors or devices for gathering and evaluating individual information. In this paper, an acoustic monitoring system for evaluation of acoustic cues is presented and tested exemplarily on four acoustic events (cough, knock, clap and phone bell). The acoustic monitoring includes several consecutive signal processing steps: the audio signal picked up by a microphone is transformed in a psycho-physiologically weighted domain using a gammatone filterbank, resulting in a so-called *cochleogram* [2]. From this cochleogram, background noise is estimated and eliminated. Further signal processing determines partitions of the cochleogram which are considered to form acoustic events. Each of these events is then evaluated with a classification algorithm. In this classification step, class membership is determined by comparing the event under inspection to representatives which have been calculated during a training phase utilizing a self-organizing map [3]. It is shown that a true positive rate from over 79% can be achieved where the false positive rate is smaller than 4% except for the event knocking.

Keywords: acoustic event detection/classification, acoustic monitoring, auditory scene analysis, self-organizing map, cochleogram.

1 Introduction

The demographic change in industrialized countries leads to an increasing amount of older people [1]. Model computations show that by 2030 half of the population in Germany will be older than 47 years [4]. Already today one eighth of the persons having accidents in the house hold which are treated in hospitals are older than 65. The risk of such accidents increases by age [4]. Nevertheless, more than 73% of the older persons want to stay in their familiar environment and live independently as long as possible [5]. Moving to a care institution is perceived as a loss in autonomy and reduction in quality of living [6].

To allow older persons to live a secure life in their own, familiar homes, the health status of older people has to be monitored, and support in every-day life has to be ensured. However, this results in a considerable effort and may be too costly to be done by nurses alone.

An automatic monitoring can support older persons as well as care personal. A monitoring device such as a video camera may be considered to be a severe break of privacy and, thus, is judged critically by the users. A possibly more accepted alternative is acoustic monitoring. If the analysis of the acoustic signal is done automatically by the system without storing acoustic data and without the possibility that other persons may listen to the signals without permission of the owner, this perhaps would not be considered as a break of privacy at all. Another advantage of acoustic monitoring is that microphones can be integrated easily and ambiently in a home.

In the following sections such an event detector/classifier based on acoustic signals and with low computational effort is presented.

In Chapter 2 of this contribution the algorithm for the detector and classifier is described. Here, a distinction will be made between the detection of an event, i.e. the case that a sound not belonging to the background noise is present, and the classification step which assigns a detected event to a previously learned class. Prior to detection and classification, a psycho-physiologically motivated pre-processing of the microphone signal is used, resulting in the so-called cochleogram. From this, background noise is removed.

Chapter 3 describes the experiments and the real environment in which the detector/classifier is evaluated. Four sounds have been selected for classification. Coughing and knocking sounds have been tested exemplarily because they may be indicators for the health status of a person: Coughing for indicating respiratory diseases and knocking for dementia.

Another tested sound is hand clapping. Clapping could be used to control and interact with technical devices as demanded by older persons, cf. [7, 8]. The developed detector/classifier is used in [9] to start a personal activity and household assistant (PAHA) which can be controlled by an automatic speech recognition (ASR) system (beneath other input and output modalities) [10].

As a fourth sound, phone ringing was tested. An acoustic detection of phone ringing could be visualized for hearing impaired persons as an additional visual alarm e.g., on a TV. More than 40% of old people report that they have difficulties in hearing such alarms from household appliances [11]. In Chapter 4 results of the detection/ classification experiments are shown and Chapter 5 concludes the paper.

2 Detector and Classifier

The acoustic detector/classifier has to be trained on example data before it can be used in the system to *learn* those events that have to be classified afterwards. In the training phase, the detector/classifier will be trained on correctly labelled example data in order to extract the model parameters of the classes which represent the specific acoustic events.

In the test phase, correctly labelled new example data are classified using the class parameters found in the training. Fig. 1 schematically illustrated the training and classification process. In both phases a continuous signal is picked up by the microphone. A pre-processing transforms the signals into a representation i.e., used by the successive detection and classification stage. The detector selects events from the continuous signal before each event is forwarded to the classifier. In the training phase a model is developed from these events. In the test phase the classifier uses the previously trained models to decide to which class a single event belongs to.

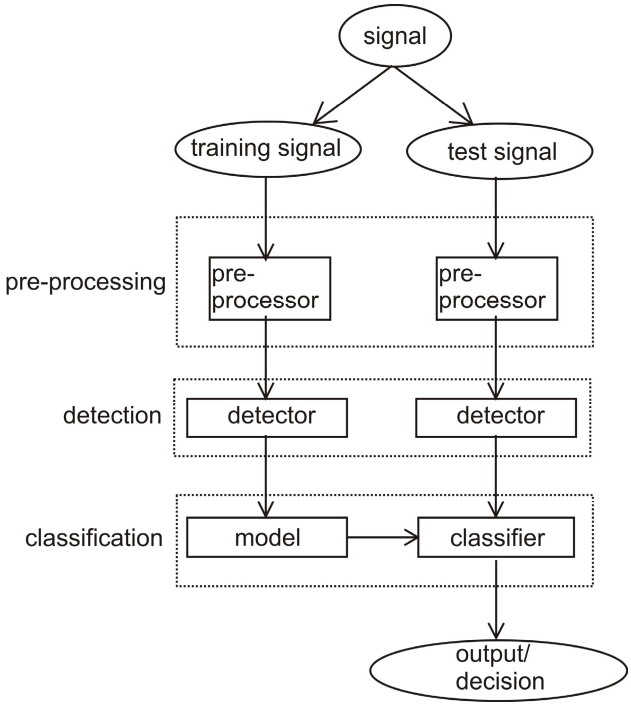

Fig. 1. Sketch of the detection and classification algorithm for training and testing.

2.1 Pre-processing

The pre-processing consists of two main steps: A transformation of the recorded time signal to a frequency-time-pattern called cochleogram and a foreground/background separation. This is done in a similar way to the aggression detector described in [2].

2.1.1 The Cochleogram as a Time-Frequency Pattern

To detect and classify an acoustic event it first has to be picked up by microphones and converted to the digital domain by an analog-digital converter. After this, it can

usually be described by a quantized amplitude over time representation. For our purpose, this might not be the best representation for a sound for several reasons. The human ear has evolutionary developed a strategy to deal with sounds which are necessary to be perceived. Inside the human ear a sound signal is amplified in the middle ear. Then, the signal is separated into overlapping frequency groups in the inner ear (cochlea). These frequency groups are not linearly distributed in the frequency domain. To account for this processing model of the human cochlea, a gammatone filterbank is used [12]. For the presented algorithm, a gammatone filterbank with $M = 93$ frequency bands is implemented. The center frequencies of the bands ranged from 20 Hz to 8 kHz in 2.85 ERB distances distributed around 1 kHz. Because the phase of a signal is irrelevant in most cases, only the logarithmic magnitude of the signal is processed. The time resolution of the filterbank output is reduced by down-sampling, i.e. each down-sampled filterbank output sample represents 5 ms.

Because the resulting frequency-time-pattern is motivated by the cochlea processing it is called cochleogram. In Fig. 2, a cochleogram of a series of coughs is shown exemplarily.

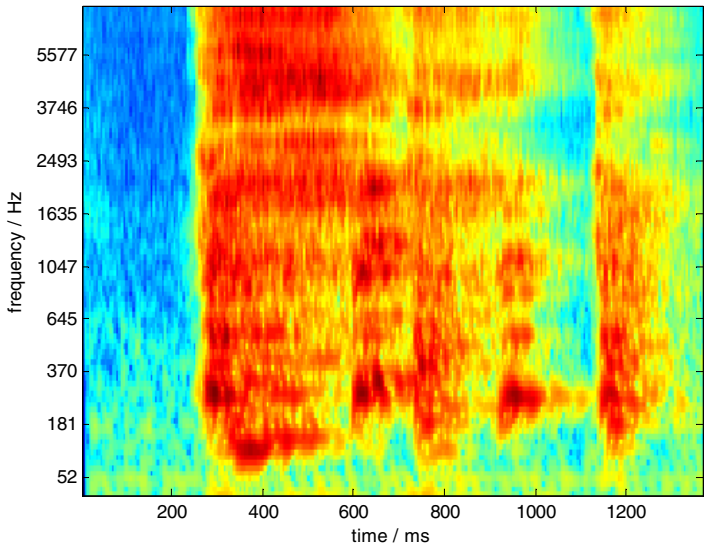

Fig. 2. Cochleogram of a series of coughs

2.1.2 Separation of Fore- and Background

A cochleogram which is denoted here by $c(m,n)$, with m being for the number of the frequency band, M the total number of bands and n the time frame, can be interpreted

as a composition of foreground sounds (to be detected) and background noise. In real-world environment which do not allow control of the kind of background noise, only very weak assumptions on frequency content or time development are possible. The only assumptions made here are that background noise level varies more slowly over time than the sounds of interest, and that no sharp frequency peaks are present in the background noise.

To separate the foreground sounds from background noise, a dynamic background model is developed which estimates the level of the background for time frame *n* by the model level of time *n*-1. This model will be described in the following.

To initialize the model, the first *I* time frames (25 frames, representing 125 ms) are averaged.

$$bg(m,0) = \frac{1}{I}\sum_{i=1}^{I} c(m,i)\,. \tag{1}$$

For all following time steps the background model at time frame *n*-1 is used as a predictor for the next time frame *n*

$$p(m,n) = \sum_{i=1}^{M} \sigma_m(i) \cdot bg(i,n-1)\ , \tag{2}$$

where $\sigma_m(i)$ is a weighting factor, that smoothes the energies of the frequency bands.

$$\sum_{i=1}^{M} \sigma_m(i) = 1 \tag{3}$$

has to be fulfilled for normalization. Usually

$$\sigma_m(i) = 0 \qquad \forall\ i < m-2 \wedge i > m+2 \tag{4}$$

holds and

$$\sigma_m(m) >> \sigma_m(i) \qquad \forall\ i \neq m\,. \tag{5}$$

Using the previous definitions, a probability mask can be generated, which weights the differences between the predicted background noise *p(m,n)* and the cochleogram *c(m,n)*.

$$\mu(m,n) = 2^{-(((c(m,n)-p(m,n))/\alpha(m))\text{^}6)} \quad . \tag{6}$$

where $\alpha(m)$ is a parameter to allow for an independent weighting of the frequency bands. This is necessary because the gammatone filters include different band widths. The mask is used to dynamically adapt the background model.

$$\begin{aligned} bg(m,n) = \quad & (1-\beta)\cdot\left[\mu(m,n)\cdot c(m,n)\right. \\ & \left.+\left(1-\mu(m,n)\right)\cdot p(m,n)\right] \\ & +\beta\cdot p(m,n) \end{aligned} \tag{7}$$

β represents the degree of dynamic adjustment. If the cochleogram *c(m,n)* is predicted well by the predictor *p(m,n)*, i.e. the difference

$$c(m,n) - p(m,n) \approx 0\,, \tag{8}$$

the mask $\mu(m,n)$ will be approximately one. In this case the new background model consists of the predictor *p(m,n)*, which is mainly the old background model *bg(m,n-1)*, and a certain amount of the current cochleogram c(m,n). If on the other hand the predictor *p(m,n)* is very different to the current cochleogram *c(m,n)*, the mask $\mu(m,n)$ will have values close to zero. This is the case when a foreground sound is dominant. The new background model *bg(m,n)* then mainly consists of the previous model *bg(m,n-1)* . In Fig. 3 the background model of the cough series from Fig. 2 is plotted.

The foreground energy is

$$f(m,n) = \left(1-\mu(m,n)\right)\cdot c(m,n)\ . \tag{9}$$

Because the background energy has been modified due to dynamic adaption, the foreground energy has to be chosen as the energy higher than the background energy

$$fg(m,n) = \begin{cases} f(m,n) & \forall f(m,n) > bg(m,n) \\ -\infty dB & otherwise \end{cases} \tag{10}$$

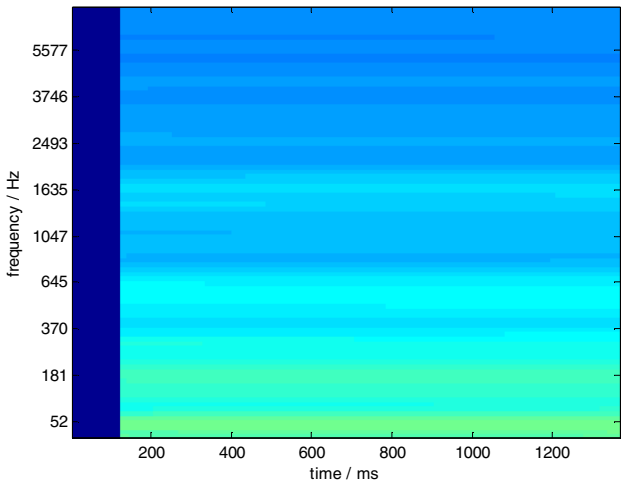

Fig. 3. Background model developed for the series of coughs from Fig. 2

The separated foreground of the cough series from Fig. 2 is plotted in Fig. 4.

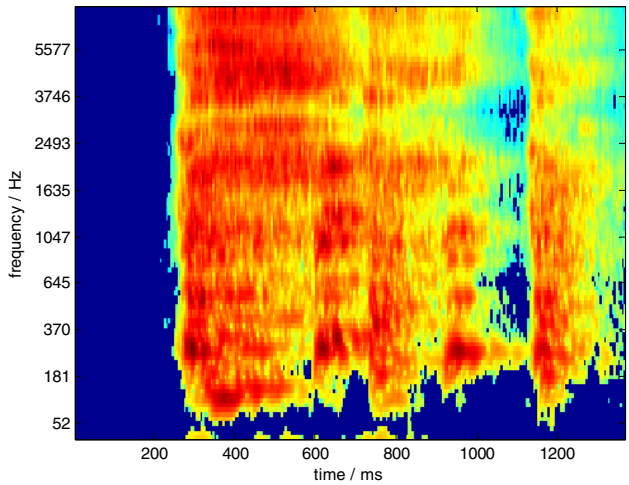

Fig. 4. Separated foreground of the cough series shown in Fig. 2

2.2 Event Detection

The detection of an event, i.e. knowledge that any sound present that does not belong to the background, is done by evaluating a threshold. If the ratio of the energy between foreground and background is higher then a defined threshold at some time

frame, this time frame is marked as start of the event. The end is marked when the energy ration drops under the threshold again.

To deal with short fluctuations of the energy around the chosen threshold, a hysteresis in time is used to determine the start of an event. The start of an event is defined when the energy ratio is larger than a chosen threshold for a pre-defined time interval which is larger than the rise time. Similar to this, the end time of an event is defined when the energy ratio falls below the threshold for a time interval larger than a defined fall time. An example is plotted in Fig. 5.

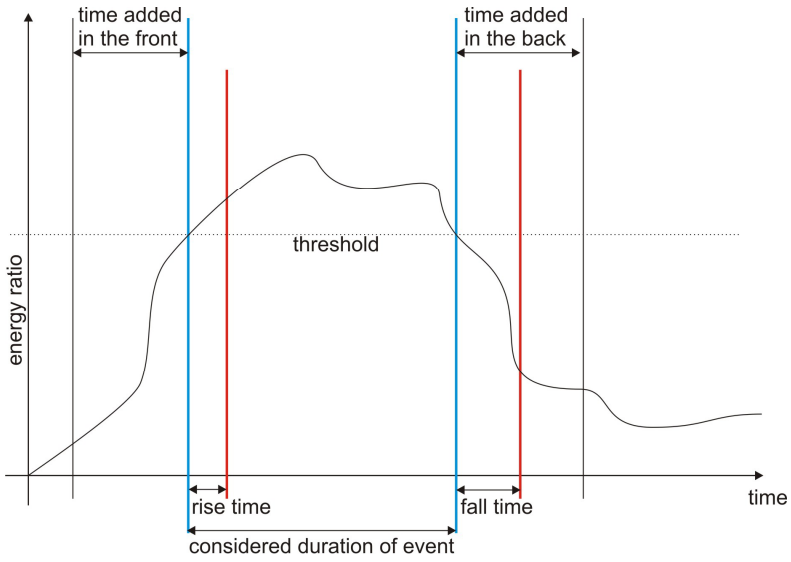

Fig. 5. Sketch of the decision step of the event detector. If the ratio between foreground and background energy stays longer than the rise time, respectively fall time over/under the threshold the start and end points of the events are marked

2.3 Classification

The two different phases for the classification process were already shown in Fig. 1: the training phase and the testing phase. During the training phase a model of the training data is developed. Afterwards, this model is used during the testing phase for decision to which class a test date belongs.

In both phases the pre-processing is the same. In this paper the separated foreground cochleograms are used. Futhermore, the means are been subtracted to achieve level independent patterns.These level independent foreground cochleograms are cut to the same length by only considering the first 180 ms (N = 36 time frames) of an event.

In the training phase each class is trained independently from the other classes. Certain amounts (about one half) of the pre-processed training data were clustered using a Self-Organizing Map (SOM) [3]. The size of the Map was 3x3. By this the

whole amount of training data was reduced to nine centroids $fg_i(m,n)$ representing the data. Centroids representing less than one date were neglected.

To evaluate a value on how much a test date differs from a centroid *i*, the normalized L1-Norm was chosen:

$$d_i = \frac{1}{M \cdot N} \sum_{m=1}^{M} \sum_{n=1}^{N} \left| fg(m,n) - fg_i(m,n) \right| . \tag{11}$$

The minimal distance was kept as a measure of difference between the test date and the class membership.

The not yet used training data were used to calculate a threshold whether the minimal distance is small enough for a positive classification, i.e. checking if the test date belongs to the tested class. The threshold was taken as the minimum that classified all training data (except some outliers with distances > 10 dB) positively.

During the testing phase only the difference between the minimal distance and the threshold has to be checked for a positive classification (member of tested class) or negative classification (not member of tested class).

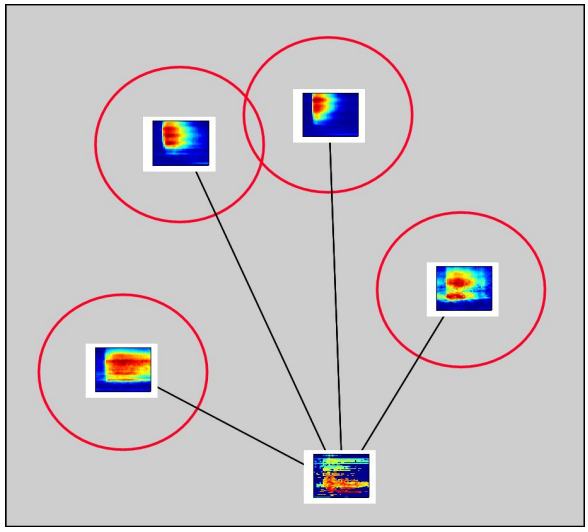

Fig. 6. Sketch of the classification algorithm. The nine centroids representing one class are placed in a high dimensional event space (gray). They are surrounded by spheres (red) representing a threshold. If a test cochleogram is inside of one of the spheres, the test date is classified positively, otherwise negatively.

In Fig. 6 a sketch represents the classification task. Nine centroids of a class (four shown here) surrounded by their threshold spheres (red) lie in the event space (gray). If the test cochleogram is inside one of these spheres the test date is classified positively otherwise negatively.

3 Experimental Setup

For evaluation of the classifier, recordings in a real environment were made. Within the Lower Saxony Research Network "Designs of Environments for Ageing" [9], a completely furnished flat has been set up where Ambient Assisted Living (AAL) technologies can be tested. For this purpose, microphones are installed in the ceiling of the living room. Different sounds produced at different positions in the room were recorded with one of the installed microphones. The recordings were made in a relatively quiet environment. The following sounds were recorded at a sampling rate of 48 kHz: coughing (two persons), hand clapping (two persons), knocking (at door, table, etc.) and phone ringing (one phone). Additionally, some every-day life recordings (conversations, bag rustling etc.) were made. The recordings of each person and each event class were labelled. Each recording session was processed by the event detector which produced the separated foreground cochleograms for each event. The detector found 33 coughs by one person and 38 by the other, 48 respectively 63 claps by both persons, 65 knocks, 9 phone rings and 36 every day life sounds.

The data were separated into training and test data to evaluate the classifier. For the two sound classes clapping and coughing, which seem to be very specific for each person, the events for one person were used for training and the remaining events for testing.

4 Evaluation

Different ways to validate a classifier are commonly used in the literature. For this paper, we chose to calculate the true and false positives and negatives.

A true positive result means that the event under test and the class on which it is tested correspond, and the classification result is correct, e.g., it is checked whether a cough signal is recognized as a cough using the cough classifier. If the classification is correct, this result is counted as a "true positive". If the event does not correspond to the classifier (a cough is evaluated with a knock classifier) and the result is negative, it is counted as a "true negative". Both classification outcomes so far have been correct. The two other possible classification outcomes are false: a cough event classified by the cough classifier to be something else would be called a "false negative"; a knock event to be classified by the cough classifier to be a cough would be called a "false positive".

In Figs. 7 to 10 the relative number of events is plotted over the minimal distances to the tested class. In Fig. 7 the classification is done on coughs, in Fig. 8 on knocks, in Fig. 9 on claps and in Fig. 10 on phone rings. The data belonging to the tested class (called test class data in the following) are displayed in black and the other data in gray. The estimated classification threshold is plotted as a vertical line. Data with minimal distances left of the threshold are classified positive and right of it negative.

In the legend the percentage of positively classified data is printed.

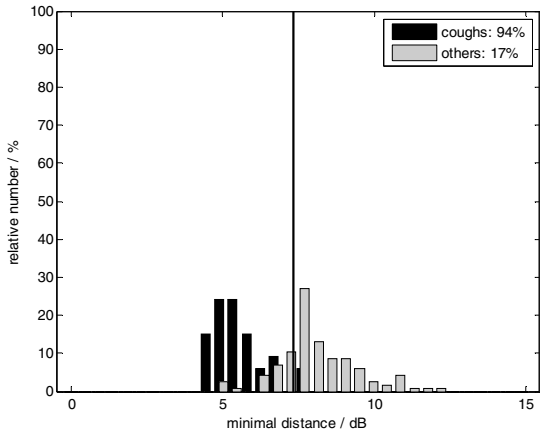

Fig. 7. Histogram of minimal distances to the cough centroids of the test class data cough (black) and the other data (gray). The vertical line shows the classification threshold at 7.3 dB. In the legend the positive rates of the data are printed in percent.

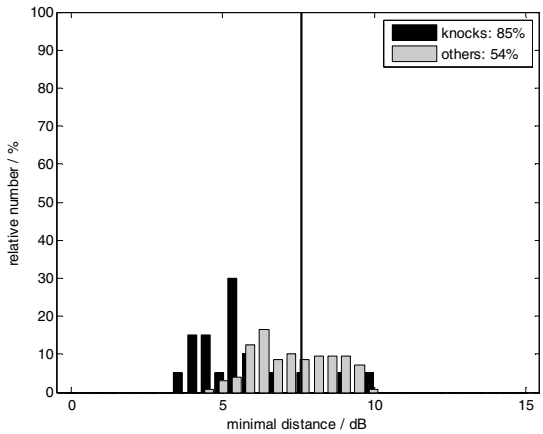

Fig. 8. Histogram of minimal distances to the knocks centroids of the test class data knocks (black) and the other data (gray). The vertical line shows the classification threshold at 7.6 dB. In the legend the positive rates of the data are printed in percent.

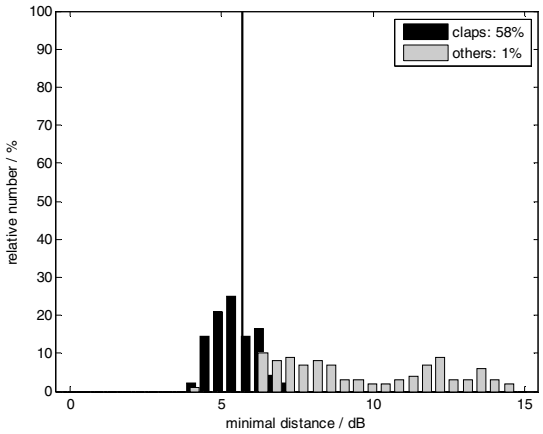

Fig. 9. Histogram of minimal distances to the clap centroids of the test class data clap (black) and the other data (gray). The vertical line shows the classification threshold at 5.7 dB. In the legend the positive rates of the data are printed in percent.

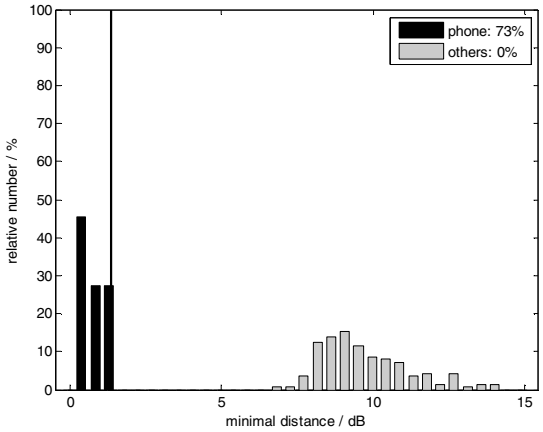

Fig. 10. Histogram of minimal distances to the phone ring centroids of the test class data phone (black) and the other data (gray). The vertical line shows the classification threshold at 1.3 dB. In the legend the positive rates of the data are printed in percent.

It can be seen that the test class data (black) in average have smaller minimal distances than the other data. A separation between these groups is possible. But the estimated classification thresholds from the training phase do not separate the classes optimal.

For phone rings the classification threshold is very low (1.3 dB) though the gap between the test class data and the other data is large enough to allow a higher threshold.

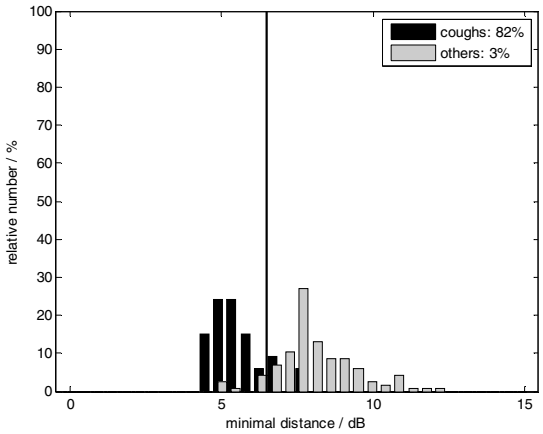

Fig. 11. Histogram of minimal distances to the cough centroids of the test class data cough (black) and the other data (gray). The vertical line shows the classification threshold at 6.5 dB. In the legend the positive rates of the data are printed in percent.

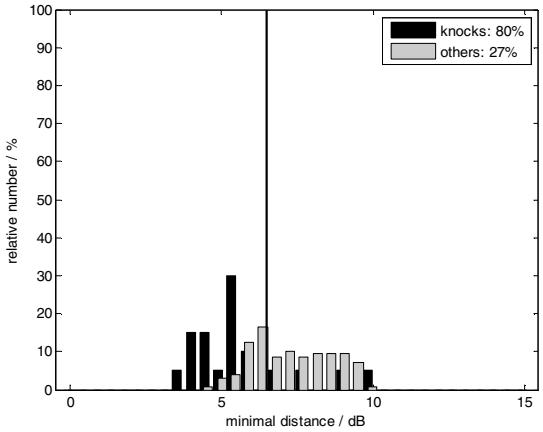

Fig. 12. Histogram of minimal distances to the knocks centroids of the test class data knocks (black) and the other data (gray). The vertical line shows the classification threshold at 6.5 dB. In the legend the positive rates of the data are printed in percent.

A fixed threshold of 6.5 dB for all experiments results in a better classification as it can be seen in Figs. 11 to 14. The true positive rate for all classes is greater than 79%, the false positive rate has been lower than 4% for all sounds except knocking. Knocking seems to have a bigger inner-class spreading due to different audio material and whether a knock was done with the open hand, the fist or the knuckles. Here, a representation by only nine centroids seems to be too few.

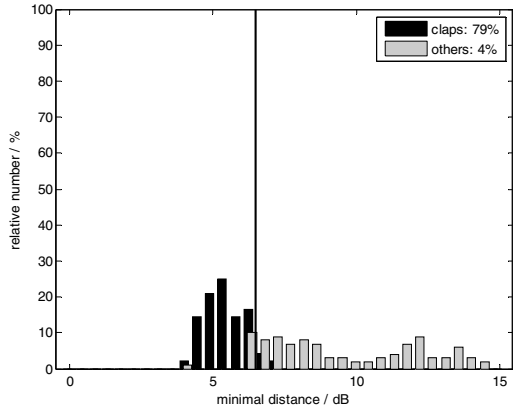

Fig. 13. Histogram of minimal distances to the clap centroids of the test class data clap (black) and the other data (gray). The vertical line shows the classification threshold at 6.5 dB. In the legend the positive rates of the data are printed in percent.

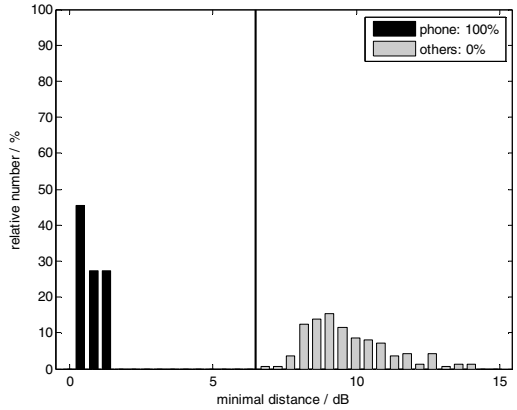

Fig. 14. Histogram of minimal distances to the phone ring centroids of the test class data phone (black) and the other data (gray). The vertical line shows the classification threshold at 6.5 dB. In the legend the positive rates of the data are printed in percent.

5 Summary

In this paper an algorithm for detection and classification of acoustic events under real conditions was presented. The pre-processing is based on the psycho-physiological motivated cochleogram. The foreground is separated from background noise. By measuring the level, a single event is marked in a continuous, acoustic input stream. The distance to representatives of a class generated from a SOM is computed, and if the minimal distance is lower than a chosen threshold, a positive classification is done.

Experiments were done in a real environment with a free field microphone installed at the ceiling of the room. It was shown that a true positive rate of 79% could be achieved for all classes where the false positive rate was lower than 4% (except for knocking).

Acknowledgements

This work was supported in parts by the Lower Saxony Ministry of Science and Culture through the "Niedersächsisches Vorab" grant programme within the Lower Saxony Research Network "Design of Environments for Ageing".

References

1. European Commision Staff. Working Document. Europes Demografic, Future: Facts and Figures. Commission of the European Communities (May 2007)
2. van Hengel, P.W.J., Andringa, T.C.: Verbal aggression detection in complex social environments. In: AVSS 2007: Proceedings of the 2007 IEEE Conference on Advanced Video and Signal Based Surveillance (2007)
3. Kohonen, T.: Self-Organizing Maps, 3rd edn. Springer, Heidelberg (2001)
4. van Hengel, P.W.J., Anemüller, J.: Audio Event Detection for In-Home Care. In: Int. Conf. on Acoustics (NAG/DAGA) (2009)
5. Grauel, J., Spellberger, A.: Akzeptanz neuer Wohntechniken für ein selbständiges Leben im Alter – Erklärung anhand soziostruktureller Merkmale. Technikkompetenz und Technikeinstellung, Zeitschrift für Sozialreform Jg 53 H 2, 191–215 (2007)
6. Schneekloth, U., Wahl, H.-W.: Selbständigkeit und Hilfsbedarf bei älteren Menschen in Privathauhalten, Pflegearrangements, Demenz, Versorgungsangebote. Kohlhammer, Stuttgart (2008)
7. Alexandersson, J., Zimmermann, G., Bund, J.: User Interfaces for AAL: How Can I Satisfy All Users? In: Proc. Ambient Assisted Living - AAL, Berlin, Germany. Deutscher Kongress, vol. 2, pp. 5–9 (2009)
8. Helal, S., Giraldo, C., Kaddoura, Y., Lee, C., El Zabadani, H., Mann, W.: Smart Phone Based Cognitive Assistant. In: Proceeding of The 2nd International Workshop on Ubiquitous Computing and Pervasive Healthcare Applications, Seattle, p. 11 (2003)
9. Project web page Lower Saxony Research Network Design of Environment for Ageing, `http://www.altersgerechte-lebenswelten.de`
10. Moritz, N., Goetze, S., Appell, J.-E.: Ambient Voice Control for a Personal Activity and Household Assistant. In: Proc. 4th German AAL-Congress, Berlin, Germany (January 2011)
11. Meis, M., Appell, J.-E., Hohmann, V., Son, N., Frowein, H., Öster, A.M., Hein, A.: Telemonitoring and Assistant System for People with Hearing Deficiencies: First Results from a User Requirement Study. In: Proceedings of European Conference on eHealth (ECEH), pp. 163–175 (2007)
12. Hohmann, V.: Frequency Analysis and Synthesis using a Gammatone Filterbank. Acta Acustica United with Acustica 88, 433–442 (2002)