Voice Activity Detection Driven Acoustic Event Classification for Monitoring in Smart Homes

Danilo Hollosi, Jens Schröder, Stefan Goetze, and Jens-E. Appell
Fraunhofer Institute for Digital Media Technology (IDMT), project group Hearing, Speech and Audio Technology (HSA)
Marie-Curie-Str. 2, 26129 Oldenburg, Germany, Email: danilo.hollosi@idmt.fraunhofer.de

Abstract—**This contribution focuses on acoustic event detection and classification for monitoring of elderly people in ambient assistive living environments such as smart homes or nursing homes. We describe an autonomous system for robust detection of acoustic events in various practically relevant acoustic situations that benefits from a voice activity detection inspired pre-processing mechanism. Therefore, various already established voice activity detection schemes have been evaluated beforehand. As a specific use case, we address coughing as an acoustic event of interest which can be interpreted as an indicator for a potentially upcoming illness. After the detection of such events using a psychoacoustically motivated spectro-temporal representation (the so-called cochleogram), we forward its output to a statistical event modeling stage for automatic instantaneous emergency classification and long-term monitoring. The parameters derived by this procedure can then be used to inform medical or care-service personal.**

I. Introduction

The continuous growth of the amount of elderly people poses great challenges to the health-care systems in many countries. This problem will become even more severe within the next years [1]. The possibility to stay in their own houses or flats independently is, thus, not only highly desired by the elderly, it will become inevitable for functioning social systems [2]. Technical systems (commonly known as ambient assistive living (AAL) technologies) are able to provide assistance to the elderly to improve daily living conditions, security and independence [2].

Various sensors can be used for monitoring different aspects of the home environment or the person's health status. Acoustic sensors in combination with appropriate signal processing strategies [3]–[7] are able to detect, analyze and track various information in smart homes unobtrusively, such as falling objects, possibly dangerous situations [8], [9] or the position of the user [10].

This work presents a three-stage approach for identification, classification and interpretation of such situations and is organized as follows. First, a concept for a system is described that combines the three-stage approach computationally (see Section II). In the next step, detailed information on each stage and its algorithmic steps are presented in Section III. This includes a description of the pre-processing stage in Section III-A, information on the event detection stage in Section III-B and our statistical approach for instantaneous emergency classification and long-term monitoring in Section III-C. Based on these information, further action can be initiated by the system, e.g., such as calling a care-service in case of a potentially dangerous event. Sections IV and V draw conclusions and present ideas for future work.

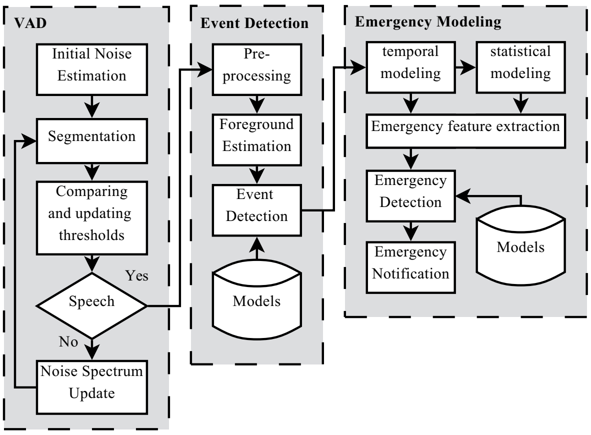

Fig. 1. System overview of the proposed system.

II. System Overview

The proposed system consists of three major processing stages: a VAD scheme to obtain low-level information about the input signal, followed by an event detection stage to derive mid-level contextual information about the data and, finally, an emergency modeling stage to formulate short- and long-time high-level semantics. A general overview of the proposed system is given in Fig. 1.

The input signal is first segmented into frames (of 32 ms length with an overlap of 27 ms) and fed to a VAD scheme that adaptively separates background noise from signal parts containing desired information. The selection of a suitable VAD scheme is based on a comparison between already established VAD algorithms. If a segment is identified to contain voice, the segment is fed to an event detector which has been initialized with a trained event model from a model database. In this work, we select a model for detecting coughs with a binary label output. These labels are then fed to an event modeling scheme to determine information about the reoccurrence, the strength and the duration of the event within a given time interval. They are the basis for a rule-based instantaneous emergency classification model and a statistical approach that allows long-term monitoring and the surveillance of the progression of an event over a longer period of time. If a potentially dangerous event is identified,

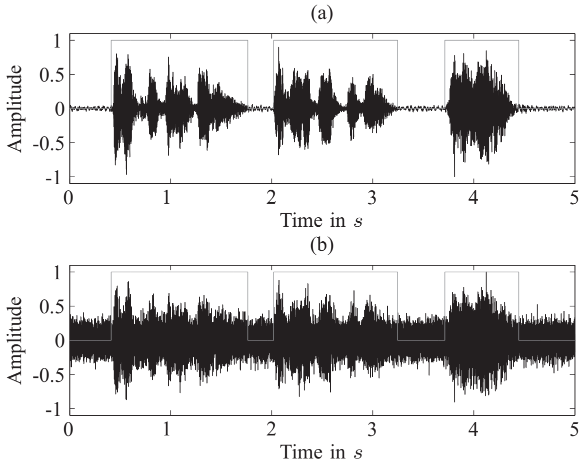

Fig. 2. VAD decision output (grey line) for an example audio recording: (a) the original recording and (b) the same recording with -15dB additive pink noise. Note that VAD becomes a non-trivial task in (b) due to the noise corruption.

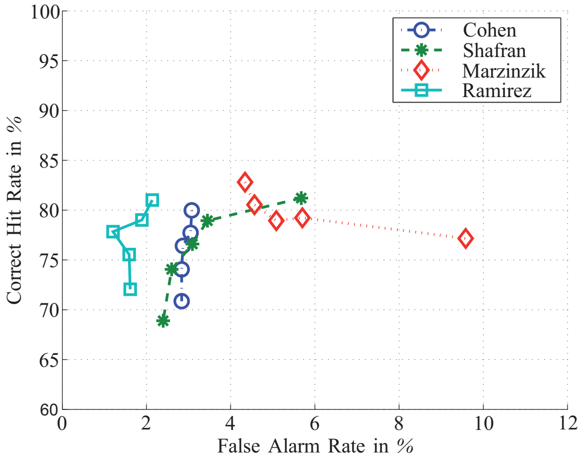

Fig. 3. The evaluation of various VAD algorithms according to their performance in noisy environments. Pink noise with a SNR in the range of 0 dB - 20 dB was added to the test data.

a message is generated to inform medical or care-service personal. The emergency model is not only limited to the detection of coughs, but can be applied to any event that is to be monitored. The proposed system works fully automatically without storing contextual information, which leads to higher end-user acceptance than e.g. video-surveillance. At the same time, it can be modified in such a way that it can be implemented in already available in-home communication systems and housing technologies. A detailed description of the system and the algorithmic steps is given in the following section.

III. Algorithm

A. Preprocessing

In general, an audio signal recorded by a microphone in real-world environment can be described as a combination of background sounds and foreground acoustic objects. Since only the latter ones are of interest they have to be separated from the background sound sources in a pre-processing stage. Therefore, an algorithmic scheme is desired which is characterized by a high sensitivity for acoustic foreground objects, a high temporal resolution to meet the event characteristics and low computational complexity.

One particular class of foreground-background separation schemes is of interest in our context: the well-studied energy-based VAD algorithms. They usually work under the assumption that an input audio signal consists of speech and stationary background noise only [21]. However, for monitoring in general and the surveillance of rooms and environments in particular, other events of non-stationary and sometimes high-leveled transient character apart from speech can be found very often. Thus, an energy-based VAD algorithm that does not include a particular speech model could classify such events as a *voice activity* as well, such that its applicability is not only limited to speech signals.

In order to select a suitable pre-processing algorithm, we evaluate several energy-based VAD schemes according to their performance under noisy conditions as depicted in Fig. 2, exemplary. In particular, we investigate the VAD schemes proposed by Marzinzik and Kollmeier [11], the long-term spectral divergence VAD method proposed by Ramirez et al. [14] and the VAD proposed by Shafran and Rose [12] as a modification of a minimum statistics based noise estimation algorithm. Inspired by this approach, we also apply a similar modification to the *Minima Controlled Recursive Averaging* (MCRA) noise estimation algorithm for speech enhancement proposed by Cohen et al. [13] due its robustness in noise and low computational costs.

The algorithms were evaluated on two hours of audio recordings captured in an office and living room environment, whereas 30 minutes of audio was used to find the optimal settings for each algorithm. All the recordings contain various noise sources such as fans, keyboard sounds, telephone rings and music. Additionally, pink noise at SNR ranging from 0 dB to 20 dB was added to the recordings to investigate the VAD performances also in highly noisy environments. In Fig. 3, the results of this experiment are illustrated as a function of *Correct Hit Rate* and *False Alarm Rate* at various SNR.

Besides a high *Correct Hit Rate*, a low *False Alarm Rate* of a suitable VAD method is desired for our application scenario to minimize the computational overhead introduced in the event classification stage. Together with the criteria mentioned at the beginning of this section, we identify the VAD method proposed by Ramirez [14] as a suitable algorithm for our work.

If an input segment is labeled as *voice activity*, it is fed to the event detection stage where the segments are further analyzed w.r.t. events of interest and potentially dangerous situations. Otherwise, the system proceeds with the next input segment.

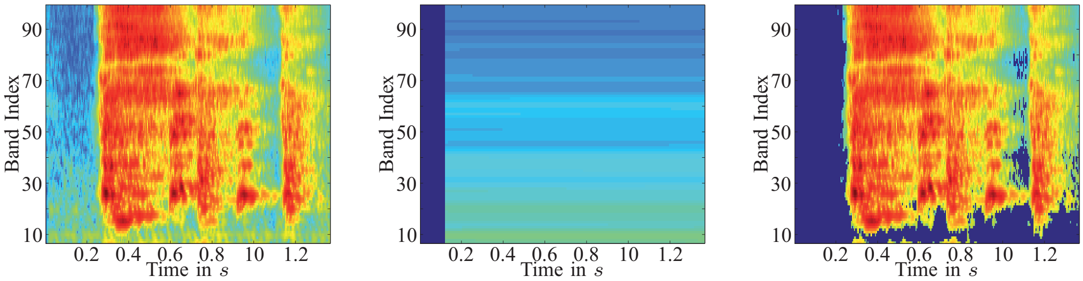

Fig. 4. Cochleogram of a series of coughs (left), its background model (middle) and the separated foreground that is used for classification (right).

B. Event Detection

In this work, we address coughing as an event of interest. At first, each segment of the input signal labeled as voice *activity* is transformed into a psychoacoustically motivated time-frequency representation in order to access both, temporal and spectral characteristics. Therefore, the so-called cochleogram [15] is calculated, which utilizes a 93 band gammatone filter-bank with center frequencies ranging from 20 Hz to 8 kHz in 2.85 ERB distances distributed around 1 kHz. The gammatone-filterbank accounts for the non-linearity in human frequency and loudness perception, such that the resulting signal representation becomes closer to the perception of sound signal within the human auditory system. At the same time, event detectors benefit from this transformation as well, as shown in [8], [9], [15].

At this processing stage, the cochleogram contains acoustic events within noisy background sounds such as e.g. fans and street noise. To avoid misclassification of the detector, the acoustic foreground objects, i.e. coughs, need to be separated from the background noises. Therefore, a dynamic background model is applied to the cochleogram that estimates the level of the background noise for a given time frame from the background noise level of previous time frames. An initial estimate for the background noise is determined by analyzing non-voice activity labeled segments in the pre-processing stage. Based on the background noise estimates, a probability mask is generated that separates the background noise from the acoustic foreground objects. Detailed information on the used foreground-background separation procedure can again be found in [15]. Please note, that the foreground-background separation can be extended to suppression of non-stationary background noises e.g. using well known approaches as in [16], [17].

An example of a coughing sequence as well as its separated background and acoustic foreground is shown in Fig. 4. Obviously, the cough sequences are characterized by highly frequently reoccurring individual events, which again underlines the necessity of a temporally fine-grained segmentation in the pre-processing stage.

The average absolute difference between the foreground and a previously trained model of the same representation is computed in the next step. In our system, we use a model for human cough detection that has been trained by analysis of various cough recordings. This step leads to a measure of the similarity between the input segment and the characteristic event pattern. If this measure becomes reasonably high, the segment is declared to contain an event that might be an indicator for a dangerous situation. Similar to the VAD algorithm described in the previous section, the event detector outputs a binary decision for each input segment that indicates whether a cough was detected or not. This information is then used to formulate statistics about the progression of the event in time and a scheme for identifying a potentially dangerous situation as it will be described in the following section.

C. Event Statistics for Emergency Classification

In this section, we describe a concept to model an emergency from its temporal characteristics. Both, short-term characteristics for instantaneous emergency classification and long-term characteristics for monitoring the progression of events over a longer period of time are covered by this concept.

Therefore, the model must reliably identify deviances between a set of parameters for an actual time interval based on the knowledge gained from previous time intervals. Since we address coughing as an event, we assume that deviant parameters correspond with a changing state of health, both in a positive and a negative way. Thus, the model should output a reliability measure to rank the current situation accordingly.

1) Instantanous Classification: For instantanous emergency classification, we continuously monitor the output of the event detector by extracting suitable parameters. An illustration of the model including the extracted parameters can be found in Fig. 5. At the beginning, the binary output of the event detector is filtered using a 1^{st} order recursive smoothing filter

$$y^*(n) = \alpha\, y(n) + (1-\alpha)\, y^*(n-1) \tag{1}$$

to obtain a suitable representation for instantaneous classification.

This step merges temporally close situated binary events, reduces the influence of sparse occurring events in the classification scheme and leads to an increase in the function $y^*(n)$ for high frequent occurring events. In the next step, a threshold T_Y is defined as the lower boundary for a potentially dangerous situation. Its duration

$$D = n_e - n_s \tag{2}$$

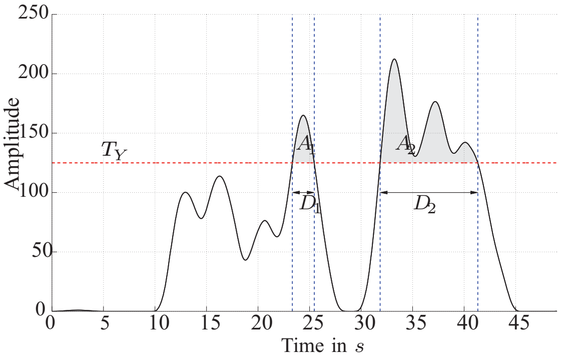

Fig. 5. Example of an Event Statistic with derived parameters

```
if $y^*(n) > T_Y$
    start calculation of $D$
    start calculation of $A$
    if $A > T_A$
        sign $emergency$
    elseif $D > T_D$
        sign $emergency$
    end
elseif $y^*(n) <= T_Y$ && $y^*(n-1) > T_Y$
    store $D$, $n_s$ and reset parameters
    store $A$, and reset parameters
end
```

Fig. 6. Pseudocode for determination of an emergency based on the extracted features duration D, acuteness A and the corresponding thresholds T_D, T_A and T_Y

is defined as the difference between the end point n_e and the start point n_s in time where the relation $y^*(n) > T_Y$ is fulfilled. An emergency is detected if the duration D of a potentially dangerous situation exceeds a predefined threshold T_D.

Additionally, we introduce the parameter acuteness

$$A = \sum_{n_s}^{n_e} y^*(n) \tag{3}$$

as a measure for the strength of a potentially dangerous situation which is calculated by integrating the pre-processed event detection function $y^*(n)$ within the duration D of a potentially dangerous situation. In our context, A is a stronger indicator for an emergency compared to parameter D.

An emergency is signed once the *Acuteness* reaches a predefined maximum T_A during integration. T_A is chosen in such that the integration time to reach the threshold is smaller compared to the duration D for high frequent reoccurring events within short periods, i.e. for very steep increases of the function $y^*(n)$. If A does not reach its predefined threshold T_A within the integration time D, an emergency can still be signed as long as D exceeds its predefined threshold T_D. The pseudocode for the instantaneous emergency classification scheme can be found in Fig. 6.

2) Long-Term Monitoring: The main idea behind the long-term monitoring model is to extract meaningful parameters from the sensor data, i.e. the audio signal captured by a microphone over a certain period of time and to put them into relation with parameters extracted from a previous time period. Therefore, a suitable time-basis needs to be defined first. Similar to previous approaches for human behavior modeling [20], we define a causal time-interval of one day, which is then sampled into a number of disjunct sub-intervals. This approach not only allows for monitoring of the event progression within sub-intervals and across them, but also along longer periods of time by forming integrally related multiples of the causal time intervals of one day. Additionally, the freedom to form even longer observation periods as a function of the desired application is guaranteed.

To analyze the progression of the events over the specified causal sub-intervals within a day, we formulate a generalized histogram as given in (4). Here, the parameters m_{t_a} and m_t denote event parameters within the actual sub-interval t_a in a causal time-interval t. The discretization parameter c is set according to the desired number of disjunct sub-intervals.

$$H(m_{t_a}) = \sum_{t=1}^{t_a} \begin{cases} 1 & \text{if } m_t - c \leq m_{t_a} \leq m_t + c \\ 0 & \text{else} \end{cases} \tag{4}$$

The choice of the sub-intervals heavily depends on the desired application and the events to monitor. At the same time, the number of sub-intervals should be chosen in such a way that widely spread distributions of the event data in the histogram are avoided. According to [18] and [19], the number of sub-intervals should not exceed 15, which otherwise would lead to insufficient capturing of reoccurring events - which is especially true for events with low a-priori probabilities - in the histogram. By taking these findings into account, a causal time interval of one day is sampled on a two hour basis, resulting in 12 sub-intervals per day.

In our scenario of human cough detection and monitoring for emergency classification, the event parameters to form long-term statistics are derived from the instantaneous emergency model described in the previous subsection. In particular, we utilize the number of events within a given sub-interval counted as the number of non-coughing to coughing transitions in $y(n)$, their temporal location in the causal time interval of one day, the temporal distance between consecutive events, i.e. coughs, the event duration D as given in (2) and the event acuteness A as defined in (3).

Additionally, we compute the mean duration and the mean acuteness of the events detected within a sub-interval as well as their variances to capture fluctuations in the event progression. In total, nine parameters are available to monitor event progressions using their histogram representations and to identify potential dangerous situations.

Once a causal time-interval is expired, the histogram can be calculated for each parameter. By comparing the histogram with data from previous time-intervals, we are now able to compute an estimation for the risk of a potential emergency. This parameter helps us to survey the health state of a person over a longer period of time based on the measures for coughing. The risk can be computed as the normalized distance of the actual event sample m_{t_a} in the histogram from normality, i.e. the maximum value of the previous histograms $H(m_t)$ as given by:

$$r(m_{t_a}) = \frac{H(m_{t_a}) - \max_{t<t_a}((H(m_t))}{\sum_t H(m_t)} \tag{5}$$

The state of health can change both in a positive and negative way. The latter one in our context is characterized by a higher number of coughs per time interval compared to the previous time intervals, a higher density of coughs around a certain point in time and a smaller temporal distance between consecutive coughs. For negatively changing health states the risk probability reaches values greater than zero, whereas for positively changing health states, the risk values stay below zero. The overall risk is defined as the weighted mean of all risks obtained from the given parameters.

$$\bar{r}(t_a) = \frac{\sum_{i=1}^{I} w_i r_i(m_{t_a})}{\sum_{i=1}^{I} w_i} \tag{6}$$

Here, w_i denotes the weight of the risk contribution of the i-th parameter for the actual time-interval. The weights are chosen according to the desired application and were set to equal $^1/_9$ in our context. The overall risk parameter $\bar{r}(t_a)$ can then be used to interpret the health state of a monitored person and to inform social and medical personal that increased attention is necessary. Furthermore, it can be used to adaptively update the thresholds in the instantaneous emergency classification stage. However, this is subject to future work.

IV. Conclusion

In this paper, we proposed a system for monitoring and emergency classification in smart homes. We showed that the combination of a pre-processing stage to obtain low-level information, an event detection stage to extract a mid-level representation from the input audio data and a stage for interpretation of this representation to high-level semantics can be used to support people in their daily life and to notify the need for intervention when necessary. Despite that a specific use case was addressed in this work, the approach presented here can be transferred to other applications and situations apart from cough detection as it is indicated by the presence of a model databases and the high degree of freedom to form temporal models in the system. Furthermore, various parameters offer the possibility to adjust the system to various environments and conditions.

V. Future Work

In order to test the system under realistic environments, extensive communication and an exchange of ideas with manufacturers, service providers, as well as health and social institutions is crucial. Replacing the rule-based approach in the instantaneous emergency classification stage with a trained classifier would allow the system to be even more flexible. Therefore, additional data is necessary for training which again underlines the necessity for further, interdisciplinary communication with industry and institutions.

References

[1] European Commision Staff, "Working Document. Europes Demografic Future: Facts and Figures," Report, May 2007.

[2] The European Ambient Assisted Living Innovation Alliance, *Ambient Assisted Living Roadmap*, VDI/VDE-IT AALIANCE Office, 2009.

[3] J. Rennies, S. Goetze, and J.-E. Appell, "Considering Hearing Deficiencies in Human-Computer Interaction," in *Human-Centered Design of E-Health Technologies: Concepts, Methods and Applications*, M. Ziefle and C. Röcker, Eds. IGI Global, 2010.

[4] S. Goetze, F. Xiong, J. Rennies, T. Rohdenburg, and J.-E. Appell, "Hands-free telecommunication for elderly persons suffering from hearing deficiencies," in Proc. *12th IEEE Int. Conf. on E-Health Networking, Appl. and Services (Healthcom'10)*, France, 2010.

[5] N. Moritz, S. Goetze, and J.-E. Appell, Ambient Voice Control for a Personal Activity and Household Assis-tant, In Proc. *4th German AAL-Kongress*, Berlin, Germany, Jan. 2011.

[6] S. Goetze, N. Moritz, J.-E. Appell, M. Meis, C. Bartsch, and J. Bitzer, "Acoustic User interfaces for ambient assisted living technologies," Accepted at *Informatics for Health and Social Care*, 2010.

[7] E. Hänsler and G. Schmidt (Eds.), *Speech and Audio Processing in Adverse Environments*, Springer, 2008.

[8] P.W.J. van Hengel and J. Anemüller, "Audio event detection for in-home care," in *Int. Conf. on Acoustics (NAG/DAGA 2009)*, 2009.

[9] P.W.J. van Hengel, M. Huisman, and J.E. Appell, "Sounds like trouble," in *Human Factors - Security and Safety*, D. de Waard, J. Godthelp, F.L. Kooi, and K.A. Brookhuis, Eds., pp. 369–375. Shaker Publishing, Maastricht, The Netherlands, 2009.

[10] T. Rohdenburg, S. Goetze, V. Hohmann, B. Kollmeier, and K.-D. Kammeyer, "Combined Source Tracking and Noise Reduction for Application in Hearing Aids," in Proc. *8. ITG-Fachtagung Sprachkommunikation*, Aachen, Germany, Oct. 2008.

[11] M. Marzinzik and B. Kollmeier, "Speech pause detection for noise spectrum estimation by tracking power envelope dynamics," in *IEEE Trans. on Speech and Audio Processing*, vol. 10, no. 2, pp. 109-118, 2002.

[12] I. Shafran, and R. Rose, "Robust speech detection and segmentation for real-time ASR applications," in Proc. *Int. Conf. on Acoustics, Speech, and Signal Processing*, (ICASSP2003),pp. 432–435, 2003.

[13] I. Cohen, and B. Berdugo "Spectral Enhancement by Tracking Speech Presence Probability in Subbands," in Proc. *IEEE Workshop on Hands Free Speech Comm.* , HSC'01, Kyoto, Japan, April 2001.

[14] J.Ramirez, J.C. Segura, C. Benitez, A. De La Torre, and A. Rubio, "Efficient voice activity detection algorithms using long-term speech information," in *Speech Communication*, vol. 42, pp. 3–4, 2004.

[15] J. Schröder and S. Wabnik and P.W.J. van Hengel, Detection und Classification of Acoustic Events for In-Home-Care, In Proc. *4th German AAL-Kongress*, Berlin, Germany, Jan. 2011.

[16] Y.Ephraim and D. Malah. "Speech enhancement using a minimum mean–square error log-spectral amplitude estimator." in *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol. 33(2), pp. 443–445, 1985.

[17] S. Goetze, V. Mildner, and K.-D. Kammeyer, "A Psychoacoustic Noise Reduction Approach for Stereo Hands-Free Systems," in Proc. *Audio Engineering Society (AES), 120th Convention*, Paris, France, May 2006.

[18] B. Gottfried, H. W. Guesgen and S. Hübner "Spatiotemporal Reasoning for Smart Homes," edited by J.C. Augusto and C.D Nugent. *Designing smart homes, Springer Verlag*, pp. 16–34, 2006.

[19] D.W. Scott "On optimal and data-based histograms." in *Biometrika*, vol. 66, pp. 605–611, 1979.

[20] G. Virone, M. Alwan, S. Dalal, S.W. Kell, B. Turner, J.A. Stankovic and R. Felder "Behavioral patterns of older adults in assisted living." in *IEEE Trans. of Information Technology in Biomedicine*, vol. 12(3), pp. 387–398, 2008.

[21] M. Wölfel and J. McDonough. "Distant Speech Recognition." *John Wiley and Sons Ltd*, 1. edition, 2009.