JCSE Journal of Computing Science and Engineering

Regular Paper

Journal of Computing Science and Engineering,
Vol. 6, No. 1, March 2012, pp. 40-50

Acoustic Monitoring and Localization for Social Care

Stefan Goetze*, Jens Schroder, Stephan Gerlach, Danilo Hollosi, and Jens-E. Appell
Fraunhofer Institute for Digital Media Technology, Project group Hearing, Speech and Audio Technology, Oldenburg, Germany
s.goetze@idmt.fraunhofer.de, jens.schroeder@idmt.fraunhofer.de, stephan.gerlach@idmt.fraunhofer.de, danilo.hollosi@idmt.fraunhofer.de, jens.appell@idmt.fraunhofer.de

Frank Wallhoff
Fraunhofer Institute for Digital Media Technology, Project group Hearing, Speech and Audio Technology, and Jade-Hochschule, Oldenburg, Germany **frank.wallhoff@idmt.fraunhofer.de**

Abstract
Increase in the number of older people due to demographic changes poses great challenges to the social healthcare systems both in the Western and as well as in the Eastern countries. Support for older people by formal care givers leads to enormous temporal and personal efforts. Therefore, one of the most important goals is to increase the efficiency and effectiveness of today's care. This can be achieved by the use of assistive technologies. These technologies are able to increase the safety of patients or to reduce the time needed for tasks that do not relate to direct interaction between the care giver and the patient. Motivated by this goal, this contribution focuses on applications of acoustic technologies to support users and care givers in ambient assisted living (AAL) scenarios. Acoustic sensors are small, unobtrusive and can be added to already existing care or living environments easily. The information gathered by the acoustic sensors can be analyzed to calculate the position of the user by localization and the context by detection and classification of acoustic events in the captured acoustic signal. By doing this, possibly dangerous situations like falls, screams or an increased amount of coughs can be detected and appropriate actions can be initialized by an intelligent autonomous system for the acoustic monitoring of older persons. The proposed system is able to reduce the false alarm rate compared to other existing and commercially available approaches that basically rely only on the acoustic level. This is due to the fact that it explicitly distinguishes between the various acoustic events and provides information on the type of emergency that has taken place. Furthermore, the position of the acoustic event can be determined as contextual information by the system that uses only the acoustic signal. By this, the position of the user is known even if she or he does not wear a localization device such as a radio-frequency identification (RFID) tag.

Category: Smart and intelligent computing

Keywords: Acoustic monitoring; Localization; Acoustic event detection and classification; Ambient assisted living

I. INTRODUCTION

The demographic changes lead to a continuous growth in the percentage of older people in today's societies [1, 2]. On the one hand, such people explicitly desire to live independently in their own homes as long as possible. On

Open Access http://dx.doi.org/10.5626/JCSE.2012.6.1.40 http://jcse.kiise.org

Received 18 July 2011, **Revised** 28 December 2011, **Accepted** 6 February 2012
*Corresponding Author

 pISSN: 1976-4677 eISSN: 2093-8020

the other hand, they have special needs due to decreasing mental and physical capabilities, such as declining strength, first mild cognitive impairment (MCI), visual or hearing impairments [3-6]. Nowadays, it is commonly accepted that the resulting problems of our care systems will not be solvable without the support of technology [7]. Examples for such assistive technologies range from reminder systems [8], medical assistance and tele-healthcare systems [9], personal emergency response systems, social robotics and safe human robot collaboration [10], to human-computer interfaces for older persons or people with special needs [8].

Assistive technology usually incorporates application dependent sensors, such as vital sensors, cameras or microphones. Recently, mobile devices have gained interest in the research community due to the availability of various sensors and communication interfaces. However, recent studies have shown that people prefer non-obtrusive sensors, such as microphones, over camera surveillance. Microphones can be easily integrated into the existing living environments in combination with appropriate signal processing strategies [6, 11-13]. It is advantageous to construct an ambient monitoring system that is able to support older persons without being noticed by the end-user. This excludes the use of wearable technology, such as smart-phones or bracelets.

We focus on the development of an acoustic monitoring system for social care applications. In nursing homes, possibly dangerous situations will be automatically detected by such systems [14-17]. In contrast to the currently available systems - which detect only an increased acoustic sound level, and using this, will trigger alarms e.g., in case of a thunder, the proposed system analyzes the signal and the context to lead to a more accurate detection of possibly dangerous situations.

Besides the ambient character of microphones in monitoring applications, they can also be used for the user interaction with technical systems. According to [8, 18], if the recognition rate is sufficiently high then, the interaction via speech or sound is natural and convenient. This kind of interaction is also preferred by older users as it was shown in recent studies [12]. A basic example is to turn on lights to enlighten a pathway for the physically impaired by just a clap of hands. With an increased computational power and deeper integration of sensors and electronic devices also more complex tasks such as giving speech commands to a technical systems, e.g., in smart homes to switch on lights, open doors or to control multimedia devices, as well as system adaptability based on the context recognition and individual user preferences becomes possible.

The remainder of this contribution is organized as follows: Section II gives an overview of the proposed system; Section III describes the incorporated signal processing strategies, i.e., signal pre-processing and analysis to obtain a robust audio signal representation, the acoustic event detection and classification stage as well as an acoustic localization and tracking of users. Together with a description of experiments the system is evaluated and its performance is discussed in Section IV and Section V concludes the paper.

Notation: Vectors and matrices are printed in boldface while scalars are printed in italic. k, n and l are the discrete time, frequency, and block index, respectively. The superscript * denotes the complex conjugation.

II. SYSTEM OVERVIEW

The system architecture adapted from [11] uses acoustic input and output for situation analysis and interaction with the user and it is schematically shown in Fig. 1. In general, the proposed system is built on a modular structure of models and signal processing strategies to serve for acoustic monitoring, emergency detection and classification and for appropriate user interaction. The system utilizes various sources of information that are either gathered before hand or they are automatically estimated during system operation. This includes the information about the acoustic environment, e.g., the position of the user, reverberation times and the damping of walls and ceilings, information about the current acoustic context, e.g. the presence and kind of noise sources, as well as the information on the individual user himself. Especially the latter is of high importance for the adjustment of the system's functionality according to the individual needs and demands of the user. Firstly, this is true for the monitoring use-case, where most of the responsibility lies on the formal care givers and secondly, for an assistance use-case, where for instance the personal audiogram of the user can be utilized to account for the individual hearing loss during human-machine interaction and cooperation [6].

The more information about the user, the environment and the context is available, the more accurate the acoustic model becomes and the better the system supports the

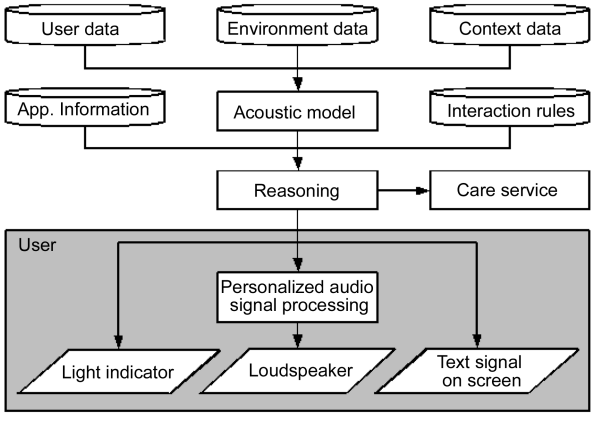


Fig. 1. Schematic structure of the emergency monitoring system.

user as well as the formal care givers. The acoustic model also yields the signal processing methods for audio pre-processing, acoustic event detection and classification, acoustic localization and tracking. They will be described in detail in the subsequent sections.

Then, a reasoning model is constructed relying on interaction rules that have been defined in close cooperation with the user as well as additional information about the desired application and use-case. This model is the core element in the system architecture. It has several input and output modalities to interact with the user or care institutions and it interprets the data according to the desired use-case. A suitable reasoning model for short- and long-term monitoring of the health status of a person and emergency detection was proposed in [16]. The detected events were treated as instantaneous mid-level representations of context under the assumption that single events do not sufficiently describe a certain situation. By taking temporal aspects and repetitions of events into account, i.e. by defining short- and long-term models for event propagation and by deriving suitable application-dependent parameters, high-level contextual information became accessible. Thus it allows for a more accurate interpretation of the situation. Due to the modular structure of the reasoning model, an adaption to other applications and use-cases is also possible.

The prediction of the reasoning model (e.g., the decision if an emergency occurred) is then transferred to the care institution or the nurses' room (e.g., as an emergency call) or to the users via an output controller. This controller manages the actual presentation of information to the user if the user lives in his/her own flat. Depending on the expected quality of the acoustic communication, the output controller initializes the information presentation in an acoustic way or by means of other modalities if the potential acoustic presentation is not suitable for the user in the given context. In this case, text messages on a screen or ambient light can be chosen for human-computer interaction.

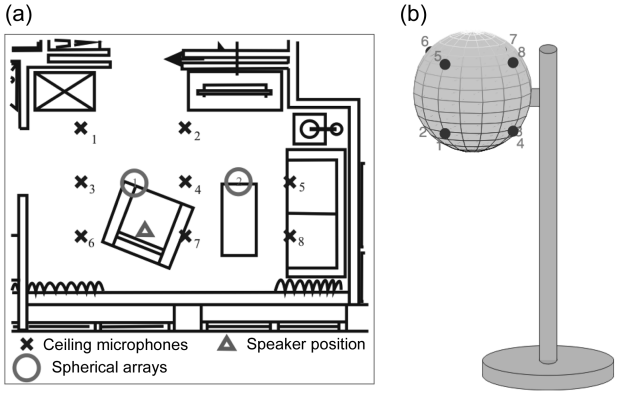


Fig. 2. (a) Positions of ceiling microphones and user positions. (b) Schematic of a spherical microphone array ambiently built in a floor lamp.

For the evaluation of the signal processing strategies described in the following an apartment was equipped with several ambient microphones in the ceiling and two spherical arrays inside a floor lamp as depicted in Fig. 2. Fig. 2a shows the positions of the ceiling microphones (crosses), of the spherical arrays (circles) as well as the position of the user (triangle). Fig. 2b shows a schematic of the 8-channel spherical microphone array built inside a floor lamp. Both the microphone arrays were used for the signal enhancement as well as position estimation of acoustic sources for the purpose of acoustic monitoring.

III. SIGNAL ANALYSIS

Humans have an astonishing ability to detect the position as well as the meaning of an acoustic source, partly relying on the information of their two ears. Therefore, the technical position detection of acoustic sources [19, 20], also relies on spatial information obtained from multiple microphones. Thus, the following subsections will present the acoustic signal processing strategies that are needed for the estimation of the position and meaning of an acoustic event for the determination of the acoustic context. The first step (cf. Section III-A) is an audio segmentation process which determines the parts of the signal that contain speech or other acoustic events. In the next step irrelevant background noise which is always present in real-world recordings is removed from the signal (cf. Section III-B). Then the signal is processed by the event detection and classification unit (cf. Section III-C and Section III-D). It determines which events are present in the signal and if a possibly dangerous situation may have occurred and whether an alarm should be raised or not. The position of the event is determined in parallel by multi-microphone position estimation and tracking system (cf. Section III-E) which serves as the context information for the detection and classification unit. Furthermore, multiple microphones allow multi-channel noise reduction schemes [21] and they are used for the separation of acoustic events from the background noise in Section III-B.

A. Audio Segmentation

For high-performance event monitoring applications, accurate signal segmentation mechanisms are of high importance. By this, the system is enabled to distinguish between the events of interest and all other acoustic signals such as, separation of speech utterances from non-speech in the acoustic stream. These methods determine the temporal location of events of interest (acoustic foreground) in a continuous audio stream and thereby, reduce the computational overhead in the later processing stages. They also provide information about the acoustic background and this is considered to be one of the strongest sources

of errors in real life monitoring applications. In order to reliably distinguish between the foreground and background acoustic objects, several methods, known as voice activity detection (VAD) algorithms have been proposed. The suitability of VAD algorithms, assume an acoustic signal to consist of solely speech and noise, for monitoring applications was evaluated in [16] using real office and living-room recordings. Following the findings that were presented, we use the long-term spectral divergence VAD method proposed by Ramirez et al. [22] in our system to obtain the labels for event presence and background noise. Then, this information can be used to initialize and trigger later processing stages as it will be described in the following subsections.

Depending on the application scenario, ethical and privacy issues may have to be considered as well. Therefore, the speech activity detection (SAD) methods can be used for audio signal segmentation [23]. In contrast to VAD algorithms, SADs contain a specific model for speech, thus they allow reliable identification of privacy related speech content in a continuous audio signal. On one hand, SADs may lead to a further increased end-user acceptance of monitoring applications due to privacy protection functionality when they are utilized in a complex system. On the other hand, for health and security related monitoring applications, SADs may corrupt the performance of the event detection and classification methods, since valuable, emergency indicating speech information (keywords, crying and shouts for help) from speech segments can not be identified any more.

B. Acoustic Foreground-Background Separation

The output of a microphone is an amplitude-time signal. Usually, this signal is corrupted by the background noise if recorded in real environments. This noise is a disturbance for the classifier and it usually decreases the classification performances of state-of-the-art classifiers, e.g. using features like Mel-frequency cepstral coefficients (MFCCs). In contrast, humans are less sensitive to the background noise disturbances [24]. Thus, in a preprocessing step, denoising algorithms are suggested [25]. In this paper, a foreground/background separation is applied that uses a probabilistic noise mask on the so-called cochleogram $c_{raw}[g, l]$ as proposed in [17]. A cochleogram is a psychophysiologically motivated spectro-temporal representation of an acoustic signal. For this purpose, the time signal from the microphone is filtered by a gammatone filterbank [26]. In this contribution, the filterbank consists of $G = 93$ gammatone filters. The center frequencies range from 20 Hz to 8,000 Hz. They are distributed in 2.85 equivalent rectangular bandwidth (ERB) distances that are around 1,000 Hz. The phase is neglected by taking the logarithm of the magnitude of the filters. The time resolution is 5 ms per frame.

For the separation of the foreground and background, a dynamically adapting background model

$$\mathbf{c}_{bg}[l] = [c_{bg}[1,l],...,c_{bg}[G,l]]^T \quad (1)$$

for every time frame l is implemented. The background model of the past is adopted to develop a prediction of the current background, $p_{bg}[g, l]$:

$$p_{bg}[g, l] = \boldsymbol{\pi}_g \cdot \mathbf{c}_{bg}[l-1], \quad (2)$$

Where, g represents a gammatone filter and

$$\boldsymbol{\pi}_g = [\pi_{g,0},...,\pi_{g,G}] \quad (3)$$

is a filter for smoothing over neighboring gammatone filter bands. For reasons of energy conservation, it has to fulfill:

$$\sum_{i=1}^{G} \pi_{g,i} = 1. \quad (4)$$

The shapes of the filters π are axially symmetrical except for the marginal g. They are only non-zero for four neighboring gammatone bands. The filterbank is plotted in Fig. 3.

A probabilistic mask is generated by the difference between the cochleogram $c_{raw}[g,l]$ as in (5) and the predicted background $p_{bg}[g, l]$,

$$\rho[g,l] = 2\text{^}-\left(\frac{c_{raw}[g,l]-p_{bg}[g,l]}{\gamma(g)}\right)^6 \quad (5)$$

Where, $\gamma(g)$ is a weighting factor to equalize the different bandwidths of the gammatone filters and $p[g,l]$ is used to extract the background:

$$c_{bg}[g,l] = (1-\beta)\cdot(p[g,l]\cdot c_{raw}[g,l] + (1-p[g,l])\cdot p[g,l]) + \beta \cdot p_{bg}[g,l], \quad (6)$$

Where, β is a factor to define the degree of influence between the current cochleogram $c_{raw}[g,l]$ and prediction $p_{bg}[g,l]$. The foreground energy due to the mask is,

$$c_{fg-mask}[g,l] = (1-p[g,l])\cdot c_{raw}[g,l] \quad (7)$$

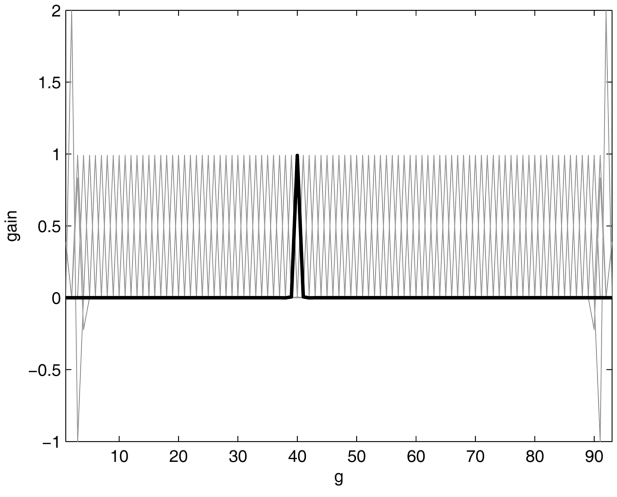


Fig. 3. The shapes of filters π_g. For better visibility, π_{40} is highlighted.

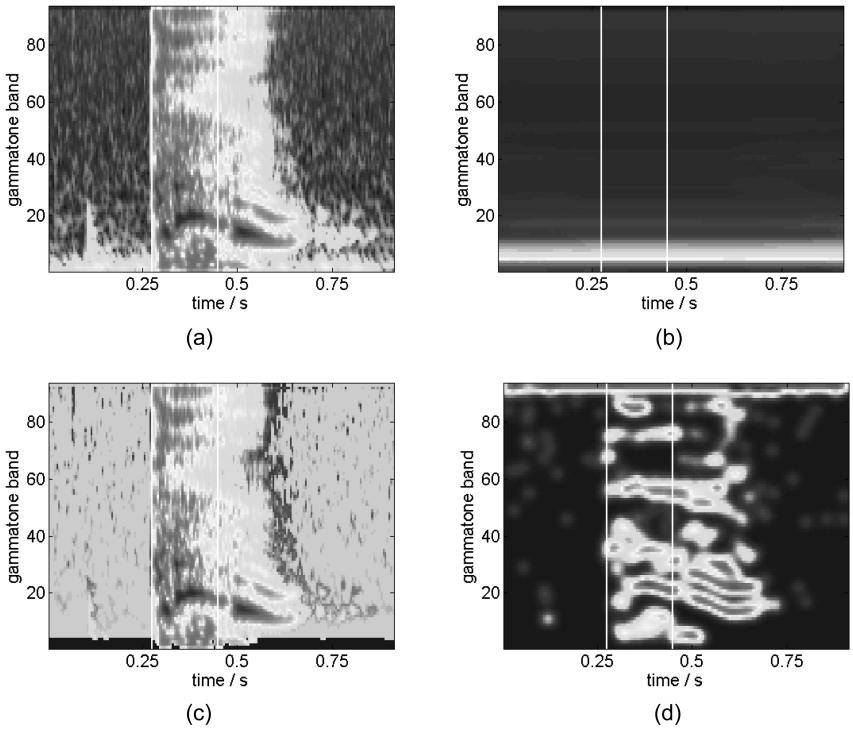


Fig. 4. Outputs of preprocessing and feature extraction stages of a human cough. (a) A raw cochleogram $\boldsymbol{c}_{\text{raw}}$. (b) A background model c_{bg}. (c) A foreground model c_{fg}. (d) A feature model $Y_{\text{v}}[\chi]$ for v = *(-1, 0)* (horizontal step) includes Gaussian spreading of edges. All panels show energy of gammatone bands over time. The first 180 ms used for the subsequent classification step are marked by white, vertical lines.

To respect the dynamic adaption of the background and to conserve the total energy, only the foreground energy that exceeds the background energy is considered:

$$c_{\text{fg}}[g,l] = \begin{cases} c_{\text{fg-mask}}[g,l] & \forall c_{\text{fg-mask}}[g,l] > c_{\text{bg}}[g,l] \\ -\infty\text{dB} & \text{otherwise.} \end{cases} \quad (8)$$

The background model has to be initialized. Hence, the first $T_{\text{init}} = 25$ frames (representing 125 ms) after the initialization of the algorithm are averaged:

$$c_{\text{bg}}[g,0] = \frac{1}{T_{\text{init}}} \sum_{l=1}^{T_{\text{init}}} c_{\text{raw}}[g,l] \quad (9)$$

In Fig. 4a-c, the raw cochleogram, the background and the foreground of a cough are plotted. It can be clearly seen that the background noise is removed from the cochleogram by the previously described method.

C. Feature Extraction

In order to extract the noise robust features, eight oriented edge detectors [27] are applied to a cochleogram. The eight orientations $\mathbf{v} = (\delta g, \delta l)$ are defined as:

$$\mathbf{v} \in \{(1,0),(1,1),(0,1),(-1,1),(-1,0),(-1,-1),(0,-1),(1,-1)\}. \quad (10)$$

Hence, edges in 45° steps are considered. The edges are detected by the derivatives Δ_{v} in the cochleograms. Instead of using only the derivatives of adjacent spectro-temporal points like the proposed in [27], smoothed derivatives over longer spectro-temporal distances are adopted here:

$$\Delta_{\text{v}}[\chi] = \sum_{i=1}^{3} c_{\text{fg}}[\chi + i \cdot \mathbf{v}] - c_{\text{fg}}[\chi - (i-1) \cdot \mathbf{v}], \quad (11)$$

In Equation (11), $\chi = (g,l)$ denotes a spectro-temporal point. A binary edge map, $Y_{\text{v}}[\chi]$ is generated by finding out the local maxima of the derivatives

$$Y_{\text{v}}[\chi] = \begin{cases} 1 & \text{if } \Delta_{\text{v}}[\chi] > \max\{\Delta_{\text{v}}[\chi - \mathbf{v}], \Delta_{\text{v}}[\chi + \mathbf{v}], \tau_\alpha\} \\ 0 & \text{otherwise} \end{cases} \quad (12)$$

Where, τ_α is a threshold of the value of the αth (= 80%) percentile of positive derivatives. In order to make $Y_{\text{v}}[\chi]$ robust to small spectro-temporal shifts, the binary edge points have to be spread. This is done by filtering $Y_{\text{v}}[\chi]$ by a two dimensional Gaussian filter with diagonal covariance matrix (standard deviation $[\sigma_g = 2,\ \sigma_l = 3]$) yielding to a feature matrix $\tilde{Y}_{\text{v}}[\chi]$. Exemplarily, this is shown for a cough in Fig. 4d.

D. Event Detection and Classification

The detection of an acoustic event, i.e. the awareness of an acoustic element that differs from the background noise, is accomplished by comparing the energy of the predicted foreground $c_{\mathrm{fg}}[g,l]$ to the background $c_{\mathrm{bg}}[g,l]$. If the ratio is higher than a defined threshold and stays above that threshold for a certain time (rise time) then, the beginning of an event is marked as depicted in Fig. 5. If the energy ratio drops again under the threshold (and stays below it for a time called as fall time) then, the end is marked. The rise and fall times are supposed to avoid the separation of a single event due to small fluctuations.

In the classification state, models for acoustic events are compared to classify the objects. Prior to this, the models have to be trained on the basis of training data. For a small amount of training data, there is a risk of over-learning models. Thus, the model complexity, i.e. the number of learned parameters, has to be kept small. Hence, an approach that can be regarded as a 1-nearest-neighbor algorithm [28] is proposed. For distance measure between a cochleogram $c_{\mathrm{fg}}[g,l]$ and a model centroid $\mu_{\mathrm{v},s,\lambda}[\chi]$ of class s, the city block distance (L1-Norm) is applied:

$$d_{s,\lambda} = \sum_{\chi,\mathrm{v}} \left| \tilde{Y}_{\mathrm{v}}[\chi] - \mu_{\mathrm{v},s,\lambda}[\chi] \right|, \tag{13}$$

Where, each class s consists of Λ_s centroids. The centroids are indexed by λ. For the difference calculation in Equation (13), $\tilde{Y}_{\mathrm{v}}[\chi]$ and $\mu_{\mathrm{v},s,\lambda}[\chi]$ must have the same dimensionality. Thus, only the first 36 frames (180 ms) of a detected event are considered (cf. white lines in Fig. 4).

The class membership s_c of $\tilde{Y}_{\mathrm{v}}[\chi]$ is estimated by finding out the centroid with the nearest distance to it:

$$s_c = \arg\min_{s} \left(\min_{\lambda} d_{s,\lambda} \right) \tag{14}$$

In order to learn centroids, k-means clustering [29] is used. Each class, s is learned separately from the others from labeled training data belonging to s. The feature $\tilde{Y}_{\mathrm{v}}[\chi]$ is extracted from every date and it is processed in the k-means algorithm. The k-means algorithm is initialized by k-means++ [30]. The cluster centers of k-means are the centroids $\mu_{\mathrm{v},s,\lambda}[\chi]$.

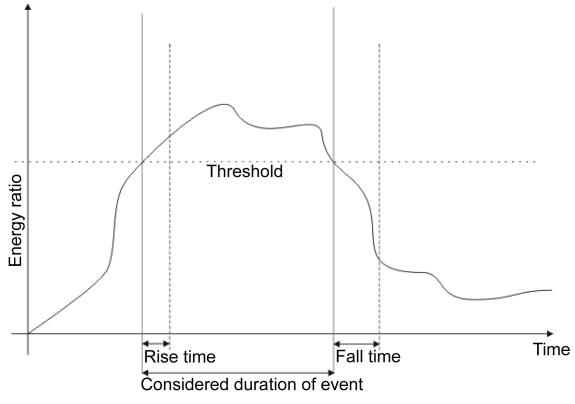


Fig. 5. Scheme of event detection. An event is marked (solid, vertical lines) by thresholding (dash-dotted, horizontal line) the energy ratio between the foreground and background (solid graph). The threshold has to be exceeded for at last the rise time respectively falls time (dashed, vertical lines).

E. Localization and Tracking of Acoustic Sources

The position of the user is important information for the reasoning and recognition system. By the exploitation of the position information it is possible, e.g., to switch on the appropriate lamp or to open the proper door by voice or sound commands, i.e. the lamp or door closest to the user. Furthermore, the position information is also highly desired if the activities of daily living (ADL) of the user should be automatically evaluated.

Acoustic position estimation is usually done by, firstly estimating the direction of arrival (DOA) of an incoming sound for several microphone pairs and secondly by combining the different DOA estimates of several microphone pairs to obtain a two-dimensional or three-dimensional position, e.g. by means of triangulation. For an overview of different DOA estimation algorithms the reader is referred to [20, 31]. For this contribution we calculate a manipulated general cross-correlation (GCC) of the microphone pairs. A spectral whitening of the input spectra is achieved by the so called phase-transform (GCC-PHAT). Regarding the characteristic of a delta-impulse one attempts to emphasize the displacement of the DOA encoding signal peak in the GCC by the phase-transform. This approach was first mentioned by Knapp and Carter [19]. The GCC-PHAT cross correlation is calculated by,

$$r_{ij}^{PHAT}[k] = \mathrm{IDFT}\{\phi_{ij}^{PHAT}[n]\}, \tag{15}$$

$$\phi_{ij}^{PHAT}[n] = \frac{\mathrm{x}_i^*[n] \cdot \mathrm{x}_j[n]}{\left|\mathrm{x}_i^*[n] \cdot \mathrm{x}_j[n]\right|}, \tag{16}$$

Where, $r_{ij}^{PHAT}[k]$ is the GCC-PHAT, $\phi_{ij}^{PHAT}[n]$ is the phase-transformed cross power spectral density (CPSD) between the microphones, i and j and $\mathrm{IDFT}\{\cdot\}$ is the inverse discrete Fourier transform. In the estimation of $\phi_{ij}^{PHAT}[n]$ in Equation (16) the phase-transform is included by the denominator $\left|\mathrm{x}_i^*[n] \cdot \mathrm{x}_j[n]\right|$. With the determination of the relevant peak position in GCC-PHAT one can calculate the DOA φ_0 by,

$$\varphi_0 = \arccos\left(\frac{\kappa_{ij} \cdot c}{f_s \cdot d_{ij}}\right). \tag{17}$$

In Equation (17), κ_{ij} denotes the peak position, f_s is the sampling frequency, d_{ij} is the distance between the microphones, and c is the sound velocity, respectively. Problems can occur during the estimation by the additional disturbing noise or interfering reverberation that results in a decrease in the detection rate.

If the system provides personalized assistance in the

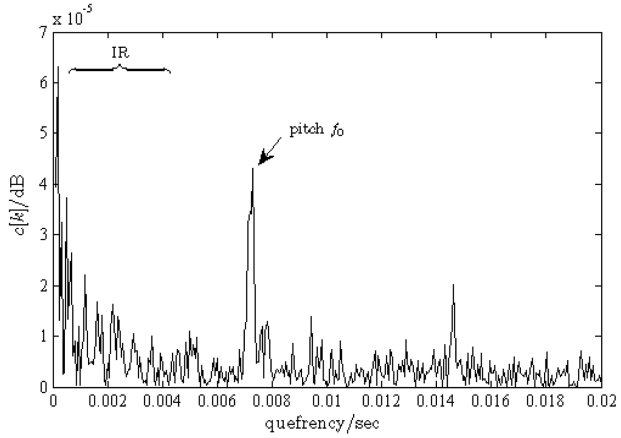


Fig. 6. Short-time cepstrum, with signal to noise ratio (SNR) = 0 dB and reverberation time T_{60} = 550 ms. The peak at 0.0072 second corresponds to a pitch of = 138 Hz (male). Preceding decline encodes the impulse response (IR) of the environment.

multiuser case then, a speaker separation is required to distinguish between the residents (e.g., male/female). The fundamental frequency (pitch) of the human voice is an important property in the human speech production to distinguish between speakers, i.e. female speakers usually have *higher* voices than male speakers. Several algorithms have been developed to estimate the pitch from the speech signals; for further details refer to the literature, e.g. [32].

In order to emphasize the pitch as a distinct peak from the disturbed input signal we use a cepstrum transformation [33]. Hence, the logarithm of the absolute value of the spectrum x[n] is inverse Fourier transformed:

$$CEP[k] = |\mathrm{IDFT}\{\log_{10}(|\mathrm{x}[n]| + \varepsilon)\}|. \qquad (18)$$

In Equation (18), $CEP[k]$ denotes the cepstrum and $\varepsilon = e^{-6}$ is a regularization factor to prevent too small values while calculating the logarithm. This transformation causes an additive representation of the signal components as depicted in Fig. 6 rather than the superimposed spectrum [33]. Thus, the so-called quefrency [34] for a dominant peak can be interpreted as a pitch. A reasonable pitch range begins at 50 Hz (low male voice) and reaches up to 500 Hz (high children voice) [35].

IV. EXPERIMENTS AND RESULTS

This section presents the experimental evaluation of the proposed monitoring system. The experimental setup for the detection and classification of the acoustic events is described in Section IV-A and the results for an acoustic localization and tracking are shown in Section IV-B.

A. Acoustic Event Detection

A database of different sounds was recorded by using a cardioid microphone with some distance from the acoustic sources. The database consisted of four classes: (hands) clapping, coughing, knocking, and phone ringing. For clapping, coughing and knocking, seven persons were asked to generate these events separated and in silence. Knocking was done on a wooden table either by using a flat hand, fist or knuckles. This decision was left to the participant. For each of these classes, 54 events were collected, where the contribution of each person differed. For phone ringing, a phone with an old fashioned ring tone was recorded nine times. Each class was separated into nine equally sized subsets to enable a nine-fold cross validation [29] and this was performed for evaluation. Three types of features were used for training and classification. It will be evaluated in the following: 1) we used the pure foreground cochleogram as described in Section III-B, 2) the smoothed edge features as described in Section III-C, and 3) standard 39-dimensional MFCCs. The window length of the MFCCs was 25 ms using 10 ms hop-time. The 0th coefficient and the derivatives of the first and second order (Δ and $\Delta\Delta$) were included to form 39-dimensional feature vectors per frame. In order to compare the results with the standard classifiers, also a Gaussian mixture model (GMM) [28, 29] using MFCC features is tested. Here, in contrast to the mentioned k-means classifier, each frame is processed separately as a feature vector. The likelihood for an event is the multiplication of the likelihoods of each frame of an event. Only the classification stage is evaluated without the investigation of the detection accuracy.

The results of the feature/classifier combinations are

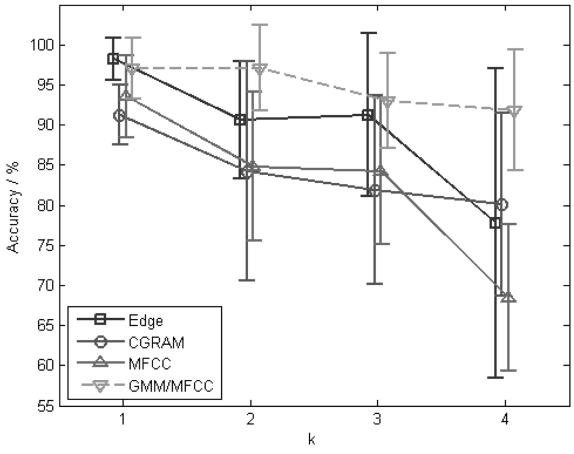
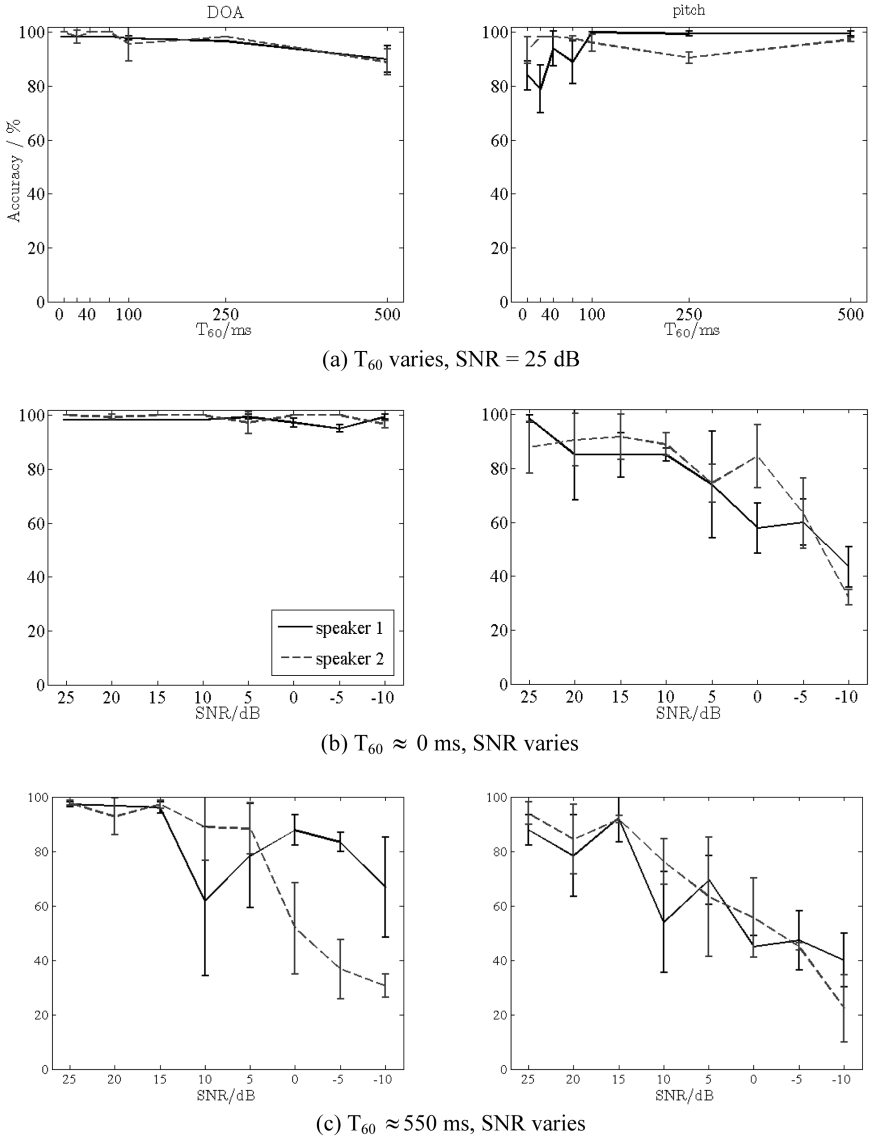


Fig. 7. Mean accuracy and standard deviation of feature/classifier combinations over the number of centroids/mixtures for the nine-fold cross validation. The feature/classifiers are the proposed Gaussian spreaded edge features with k-means (Edge, squares), pure foreground cochleograms (CGRAM) with k-means (circles), Mel-frequency cepstral coefficients (MFCCs) with k-means (MFCC, upward-pointing triangles) and MFCCs with Gaussian mixture model (GMM) (GMM/MFCC, downward-pointing triangles, dashed line).

shown in Fig. 7. The mean accuracies and standard deviations of the nine trials of the cross validation are plotted over a number of centroids/mixture components Λ ($\Lambda_s = \Lambda \forall s$). The accuracies are best for $\Lambda = 1$. This is probably an effect of the small training database. If more centroids are generated by the k-means then, it becomes more sensitive to the outliers. Only for $\Lambda = 1$, stable results with standard deviation less than 5% are calculated, i.e. for a small database the k-means algorithm can be replaced by just calculating the mean of the features. The GMM does not show this high dependency on the number of means as it processes the frames separately. Instead of the maximal 48 features per class as for the k-means, the GMM can revert to over 2,000 feature vectors (for phone ringing less). For the case $\Lambda = 1$, the edge features outperform the others. Even the GMM/MFCC combination is worth, but it is only 1%. Moreover, it is insignificant for this small database. The performance of other k-means classified features lie more than 5% under the edge based k-means classifier. In summary, Fig. 7 shows that a recognition rate of >90% can be achieved by GMM based approaches and for $\Lambda < 4$ which also uses the computationally extremely simple Edge feature. Thus, a classification of everyday acoustic events is possible even under realistic acoustic conditions.

B. Localization and Tracking

For the evaluation of the localization sub-system, a combined pitch and DOA estimation algorithm is used



(a) T_{60} varies, SNR = 25 dB

(b) $T_{60} \approx 0$ ms, SNR varies

(c) $T_{60} \approx 550$ ms, SNR varies

Fig. 8. Accuracy rates and standard deviation for two concurrent speaker (male/female) in terms of direction of arrival (DOA, left) and pitch (right) estimation. SNR: signal to noise ratio.

based on a harmonic sieve filtering of the CPSD that includes the concept of the cepstrum and GCC-PHAT calculation [36]. Recently, the improvements to this combined algorithm were proposed [37]. The setup was chosen to correspond to the room that is depicted in Fig. 2. Eight signal to noise ratios (SNRs) that ranges from 25 dB to -10 dB and eight reverberation times that ranges from 0 ms (no reverberation) to 500 ms (corresponding to common office environment) were simulated. An SNR of 25 dB implies that there was hardly any noise that was perceived by human listeners and at an SNR of -10 dB hardly any signal can be perceived (e.g. as in a quickly driving car). Furthermore, real reverberation measurements where used for additional evaluations. SNR is adjusted by additional spectrally speech shaped, diffuse noise that is uncorrelated to the desired signal. A six channel line array with 22 cm distance between each microphone was adopted. As a valuation standard the accuracy rate of all the estimations was calculated, tolerating a deviation of 10^{o} and 10 Hz, respectively, i.e. if the direction estimate was within a tolerance of ±10 degrees and the pitch estimate to distinguish between the speakers was within a tolerance of ±10 Hz then, the estimate was considered to be correct.

Two concurrent speakers (male/female) were present and they had to be tracked.

In Fig. 8 the results for different scenarios are depicted. Fig. 8a show the accuracy rates for the different reverberation times T_{60} for a clean speech. The results are quite good without any exception for any room reverberation time and for both speakers as well for the direction estimate (left panels) as for the pitch estimate (right panels). Results for a scenario with varying SNR but without reverberation are shown in Fig. 8b. Especially the pitch estimation declines with lower SNR down to 40% for -10 dB SNR. However, it should be noted that common SNR which is to be expected in a household or a care environment is >10 dB. Fig. 8c again shows the results for the different SNR values. This time, real measured room impulse responses (RIRs) with reverberation of about $T_{60} \approx 550$ ms instead of simulated ones are used. These data show a particular impact to the DOA estimation when compared to the Fig. 8b. The results show that the reliable results can be expected even under moderately high reverberation up to the noise levels of 5-10 dB SNR (~70% accuracy) which is sufficient for the envisaged household and care scenarios.

V. CONCLUSION

An acoustic monitoring system for an ambient assisted living scenario includes technologies that are to determine the user's activity by means of an acoustic event detection and classification. Localization of the user was also described in this contribution. The system combined strategies for signal segmentation, background noise reduction, event detection and classification as well as acoustic source position estimation and tracking to lead to a practically applicable overall system. The system was evaluated under acoustically realistic conditions including disturbances such as ambient noise and reverberation. It was shown that as well the position of the user as contextual information for the emergency monitoring system as well as acoustic events can be reliably detected.

ACKNOWLEDGMENTS

This work was supported in parts by the German Ministry of Education and Research (BMBF) under grant no. V4KMU10/117 and the German Lower Saxony Ministry of Science and Culture through the "Niedersachsisches Vorab" grant programme (project GAL).

REFERENCES

1. Commission of the European Communities, "Europe's demographic future: facts and figures," Staff Working Document SEC (2007) 638, Brussels: Commission of the European Communities, 2007.
2. Statistisches Bundesamt [Federal Statistical Office], "Demografischer wandel in Deutschland: auswirkungen auf krankenhausbehandlungen und pflegebedürftige im bund und in den landern [Demographic change in Germany: impacts on hospital treatments and people in need of care]," 2008, Available: https://www.statistik.bayern.de/veroeffentlichungen/download/A1832E%20200851/A1832E%20200851.pdf. German.
3. R. C. Petersen, "Mild cognitive impairment as a diagnostic entity," *Journal of Internal Medicine*, vol. 256, no. 3, pp. 183-194, 2004.
4. M. A. Rudberg, S. E. Furner, J. E. Dunn, and C. K. Cassel, "The relationship of visual and hearing impairments to disability: an analysis using the longitudinal study of aging," *Journal of Gerontology*, vol. 48, no. 6, pp. M261-M265, 1993.
5. S. Uimonen, K. Huttunen., K. Jounio-Ervasti, and M. Sorri, "Do we know the real need for hearing rehabilitation at the population level? Hearing impairments in the 5- to 75-year-old cross-sectional Finnish population," *British Journal of Audiology*, vol. 33, no. 1, pp. 53-59, 1999.
6. S. Goetze, F. Xiong, J. Rennies, T. Rohdenburg, and J. E. Appell, "Hands-free telecommunication for elderly persons suffering from hearing deficiencies," *12th IEEE International Conference on E-Health Networking, Application and Services*, Lyon, France, 2010, pp. 209-224.
7. G. van den Broek, F. Cavallo, and C. Wehrmann, AALIANCE Ambient Assisted Living Roadmap, Amsterdam, Netherlands: IOS Press, 2010.
8. S. Boll, W. Heuten, E. M. Meyer, and M. Meis, "Development of a multimodal reminder system for older persons in their residential home," *Informatics for Health and Social Care*, vol. 35, no. 3-4, pp. 104-124, 2010.
9. C. Lisetti, F. Nasoz, C. LeRouge, O. Ozyer, and K. Alvarez, "Developing multimodal intelligent affective interfaces for

tele-home health care," *International Journal of Human-Computer Studies,* vol. 59, no. 1-2, pp. 245-255, 2003.
10. S. Chew, W. Tay, D. Smit, and C. Bartneck, "Do social robots walk or roll?," *Proceedings of the Second International Conference on Social Robotics*, Singapore, 2010, pp. 355-361.
11. J. Rennies, S. Goetze, and J. E. Appell, "Personalized acoustic interfaces for human-computer interaction," *Human-Centered Design of E-Health Technologies: Concepts, Methods and Applications*, Hershey, PA: IGI Global, 2011. pp. 180-207.
12. S. Goetze, N. Moritz, J. E. Appell, M. Meis, C. Bartsch, and J. Bitzer, "Acoustic user interfaces for ambient-assisted living technologies," *Informatics for Health and Social Care*, vol. 35, no. 3-4, pp. 125-143, 2010.
13. E. Hansler and G. Schmidt, Speech and Audio Processing in Adverse Environments, Heidelberg, Germany: Springer, 2008.
14. P. W. J. van Hengel and J. Anemuller, "Audio event detection for in-home care," *NAG/DAGA International Conference on Acoustics*, Rotterdam, Amsterdam, 2009, pp. 618-620.
15. P. W. J. van Hengel, M. Huisman, and J. E. Appell, "Sounds like trouble," *Human Factors Security and Safety*, Maastricht, Netherlands: Shaker Publishing, 2009, pp. 369-375.
16. D. Hollosi, J. Schroder, S. Goetze, and J. E. Appell, "Voice activity detection driven acoustic event classification for monitoring in smart homes," *3rd International Symposium on Applied Sciences in Biomedical and Communication Technologies*, Rome, Italy, 2010, pp. 1-5.
17. J. Schroder, S. Wabnik, P. W. J. van Hengel, and S. Goetze, "Detection and classification of acoustic events for in-home care," Ambient Assisted Living, Heidelberg, Germany: Springer, 2011, pp. 181-196.
18. B. Pfister and T. Kaufmann, "Sprachverarbeitung: grundlagen und methoden der sprachsynthese und spracherkennung," Heidelberg, Germany: Springer, 2008.
19. C. H. Knapp and G. C. Carter, "The generalized correlation method for estimation of time delay," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 24, no. 4, pp. 320-327, 1976.
20. G. Doblinger, "Localization and tracking of acoustical sources," *Topics in Acoustic Echo and Noise Control*, Heidelberg, Germany: Springer, 2006, pp. 91-122.
21. J. Bitzer and K. U. Simmer, "Superdirective microphone arrays," *Microphone Arrays: Signal Processing Techniques and Applications*, Heidelberg, Germany: Springer, 2001, pp. 19-38.
22. J. Ramirez, J. C. Segura, C. Benitez, A. de la Torre, and A. Rubio, "Efficient voice activity detection algorithms using long-term speech information," *Speech Communication*, vol. 42, pp. 3-4, pp. 271-287, 2004.
23. N. Cho and E Kim, "Enhanced voice activity detection using acoustic event detection and classification," *IEEE Transactions on Consumer Electronics*, vol. 57, no. 1, pp. 196-202, 2011.
24. B. T. Meyer, T. Brand, and B. Kollmeier, "Effect of speech-intrinsic variations on human and automatic recognition of spoken phonemes," *Journal of the Acoustical Society of America*, vol. 129, no. 1, pp. 388-403, 2011.
25. D. Macho, L. Mauuary, B. Noe, Y. M. Cheng, D. Ealey, D. Jouvet, H. Kelleher, D. Pearce, and F. Saadoun, "Evaluation of a noise-robust DSR front-end on aurora databases," *7th International Conference on Spoken Language Processing*, Denver, CA, 2002, pp. 17-22.
26. V. Hohmann, "Frequency analysis and synthesis using a Gammatone filterbank," *Acta Acustica united with Acoustica*, vol. 88, no. 3, pp. 433-442, 2002.
27. Y. Amit, A. Koloydenko, and P. Niyogi, "Robust acoustic object detection," *Journal of the Acoustical Society of America*, vol. 118, no. 4, pp. 2634-2648, 2005.
28. R.O. Duda, P. E. Hart, and D. G. Stork, Pattern Classification, 2nd ed. New York, NY: Wiley, 2001.
29. C. M. Bishop, Pattern Recognition and Machine Learning, Heidelberg, Germany: Springer, 2006.
30. D. Arthur and S. Vassilvitskii, "k-means++: the advantages of careful seeding," Stanford InfoLab, Stanford, CA, Technical Report 2006-13, 2006.
31. J. Chen, J. Benesty, and Y. A. Huang, "Time delay estimation in room acoustic environments: an overview," *EURASIP Journal on Applied Signal Processing*, vol. 2006, pp. 1-19, 2006.
32. A. Savard, "Overview of homophonic pitch detection algorithms," Schulich School of Music, McGill University, Montreal, Technical Report MUMT-612, 2006.
33. A. M. Noll, "Cepstrum pitch determination," *Journal of the Acoustical Society of America*, vol. 41, no. 2, pp. 293-309, 1967.
34. A. V. Oppenheim and R. W. Schafer, "From frequency to quefrency: a history of the cepstrum," *IEEE Signal Processing Magazine*, vol. 21, no. 5, pp. 95-106, 2004.
35. P. Vary and R. Martin, Digital Speech Transmission: Enhancement, Coding and Error Concealment, Chichester, UK: John Wiley & Sons Ltd, 2006.
36. M. Kepesi, F. Pernkopf, and M. Wohlmayr, "Joint position-pitch tracking for 2-channel audio," *5th International Work-*

Stefan Goetze

Stefan Goetze is head of Audio System Technology for Assistive Systems at the Fraunhofer Institute for Digital Media Technology (IDMT), project group Hearing, Speech and Audio (HSA) in Oldenburg, Germany. He received his Dipl.-Ing. in 2004 at the University of Bremen, Germany, where he worked as a research engineer from 2004 to 2008. His research interests are assistive technologies, sound pick/up and enhancement as well as detection and classification of acoustic events and automatic speech recognition. He is lecturer at the University of Bremen and project leader of national and international projects in the field of ambient assisted living (AAL).

Jens Schroder

Jens Schroder is a PhD candidate at the Fraunhofer Institute for Digital Media Technology (IDMT) in Oldenburg, Germany. He graduated in physics at the Carl-von-Ossietzky University of Oldenburg, Germany, in 2009. His research interests include statistical modeling, pattern recognition and classification of acoustic events.

Stephan Gerlach

Stephan Gerlach received his Dipl.-Ing. (FH) in Electrical Engineering from the University of Applied Sciences Magdeburg-Stendal (FH), Germany in 2008. In 2011 he received his M.Sc. in Hearing Technology and Audiology from the Carl-von-Ossietzky University of Oldenburg. Currently Stephan Gerlach is a PhD candidate at the Fraunhofer Institute for Digital Media Technology (IDMT) in Oldenburg, Germany. His research interests include audio signal processing, especially source localization in acoustic sensor networks.

Danilo Hollosi

Danilo Hollosi recieved his Diploma degree in Mediatechnology -Audiovisual Technology from the Technical University of Ilmenau, Germany, in 2009. He currently holds a research position at the Fraunhofer IDMT-HSA in Oldenburg, Germany. His research interests are audio signal processing, machine learning and modeling and music information retrieval.

Jens-E. Appell

Jens-E. Appell received the physics diploma (Diplom-Physik) degree from University of Gottingen and Ph.D. degree from the Carl-von-Ossietzky University of Oldenburg, Germany, in 1994 and 2002. From 2001 until 2008 he was director of the embedded HW/SW division and head of the Design Center with a research focus on embedded system design and AAL applications at OFFIS - Institute for Information Technologie, Germany. Amongst several other projects in FP5 and FP6 he was the coordinator of the FP6 project hearing@home. Since 2008 he is the head of department for Hearing, Speech and Audio Technology of the Fraunhofer Institute for Digital Media Technology, Germany.

Frank Wallhoff

Frank Wallhoff studied Electrical and Information Engineering at Duisburg University. In 2006 he received the Dr.-Ing. degree at Technische Universitat Munchen (TUM), Munich, Germany, where he initiated the Interactive Systems Research Group at the Institute for Human-Machine Communication within the Cluster of Excellence CoTeSys. In 2010, he became Professor for Assistive Technologies at the Jade University of Applied Sciences in Oldenburg. Besides other projects, Prof. Dr.-Ing. Frank Wallhoff is coordinator of the FP7 project Custom Packer and the AAL-Joint-Programme project ALIAS both with focus on robotics.

shop on Content-Based Multimedia Indexing, Bordeaux, France, 2007, pp. 303-306.