Online Diarization of Streaming Audio-Visual Data for Smart Environments

Joerg Schmalenstroeer and Reinhold Haeb-Umbach, *Senior Member, IEEE*

***Abstract*—For an environment to be perceived as being smart, contextual information has to be gathered to adapt the system's behavior and its interface towards the user. Being a rich source of context information speech can be acquired unobtrusively by microphone arrays and then processed to extract information about the user and his environment. In this paper, a system for joint temporal segmentation, speaker localization, and identification is presented, which is supported by face identification from video data obtained from a steerable camera. Special attention is paid to latency aspects and online processing capabilities, as they are important for the application under investigation, namely ambient communication. It describes the vision of terminal-less, session-less and multi-modal telecommunication with remote partners, where the user can move freely within his home while the communication follows him. The speaker diarization serves as a context source, which has been integrated in a service-oriented middleware architecture and provided to the application to select the most appropriate I/O device and to steer the camera towards the speaker during ambient communication.**

***Index Terms*—ambient communication, diarization, middleware.**

I. Introduction

AMBIENT intelligence (AmI) describes the vision of technology that is invisible, embedded in our surroundings and present whenever we need it. Interacting with it should be simple and effortless, or, as E. Aarts puts it, "*The systems can think on their own and can make our lives easier with subtle or no direction*" [1].

From the early days of this computing and interaction paradigm, speech has been considered a major building block of AmI [2]. The purpose of speech and audio processing is twofold:

- speech as an input/output modality that facilitates a simple and effortless interaction with the system, preferably in cooperation with other modalities;
- speech and other acoustic events as a source of context information.

In this paper, we are concerned with the latter, where the term *context* has to be understood very broadly as any piece of information that may be relevant to the system, such as information about the (physical) environment, the computing equipment or the user.

Context information is of paramount importance for an AmI system. Only if a system is able to adapt to the environment, the user and the task in an unobtrusive, seamless manner it will be perceived as intelligent. An AmI system should therefore be able to extract context information from sensor signals, merge different pieces of information and reason about it, all without explicit assistance by the user.

Here we deal with context acquisition from audio or audio-visual signals. We all know from our interpersonal communication experience that speech is a very rich source of context information. It not only tells us about the "what" (the contents), but also about "who speaks where and when, and even how," since the speech signal conveys information about the speaker's identity, location, and emotion.

The extraction of information from speech, other than by recognition and understanding, has attracted increased interest by the research community in recent years. In the "rich transcription tasks" sponsored by DARPA the goal was a speaker diarization of broadcast news, meeting recordings or telephone conversations [3] in order to attribute temporal regions to specific speakers. In these applications, usually a batch processing scenario is considered, i.e., the complete recording of the audio data is available at the beginning of the diarization process. For such a setup iterative or multiple-pass procedures have been devised which go over the data many times [4]–[6].

Context acquisition from audio or audio-visual sensor signals for an AmI system poses some unique challenges, which are quite different from the offline scenario studied in the rich transcription tasks. First of all, an online processing has to be conducted, where an audio stream is segmented and classified "on the fly" with as little latency as possible. New context information should be available to the AmI system as quickly as possible such that it can react upon it, e.g., by adapting its interface to the user and environment [7]. Relatively few publications on diarization are concerned with this online aspect [8]–[10]. Another distinctive difference to many speaker diarization tasks is that we are also interested in the position of the speaker in the room, as this can be used for example to select the most appropriate input/output (I/O) device with respect to the location of the user. We therefore employ microphone array signal processing, from which the speaker position can be obtained. The position information has also turned out to be very helpful for improving the audio segmentation and speaker identification accuracy [11], [12].

The specific application we are considering here is ambient communication or ambient telephony [13], [14]. These terms have been phrased to describe telecommunication with a re-

Manuscript received December 09, 2008; revised March 30, 2010; accepted May 05, 2010. Date of publication May 18, 2010; date of current version September 15, 2010. This work was supported in part by the European Union project "Amigo—Ambient Intelligence for the Networked Home Environment" under Contract IST 004182. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. James Glass.

The authors are with the Department of Communications Engineering, University of Paderborn, 33095 Paderborn, Germany (e-mail: schmalen@nt.uni-paderborn.de; haeb@uni-paderborn.de).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/JSTSP.2010.2050519

mote partner, which can be carried out in the same way as if the partner were physically present, i.e., the communication is terminal-less (terminals are hidden in the environment; a user is allowed to move freely within his home, while the communication follows him), session-less (the line is open for a long time and communication is carried out at varying levels of intensity) and multi-modal (preferably audio-visual). For this application the audio-visual processing serves two purposes: First, the diarization results are used to adapt the user interface to the context (selection of most appropriate I/O-device, loading of presets for the identified user), and, second, the location information gleaned from the audio-visual input is used to control the camera.

The challenges faced with this application share some commonalities with the issues studied within the European Union-sponsored projects Computer in the human interaction loop [15] (CHIL) and Augmented Multi-Party Interaction [16] (AMI). Their aim was to improve human interaction and business productivity by advanced signal processing, and, among others, important work on speaker localization and identification, as well as on acoustic event detection has been carried out [9], [17]. However, the projects focused on different applications, such as lecture and meeting transcription, with obviously different constraints and objectives than ambient communication. Also the setup was in some part different, e.g., in AMI multiple microphones were used but typically not preconfigured microphone arrays, so position information could not be extracted [18].

Other work on audio-visual scene analysis has been and is currently under way in the fields of advanced video-conferencing systems, automatic surveillance systems and systems for ambient assisted living, see, e.g., [19], [20]. For these applications, a lot of work has been devoted to robust audio-video localization, and it has been shown that, if the varying reliability of the audio and visual input data is taken into account, robust tracking is achieved even if one of the modalities is temporarily of low quality [21].

The focus of the work described in this paper, however, is not primarily on robustness, as we assume a cooperative user who is willing to communicate and thus faces the camera most of the time. We also assume that the users have gone through an enrollment phase such that speaker and face models have been trained prior to usage. While these assumptions clearly indicate some limitations of our work, the remaining problem of diarization of streaming audio-visual data studied here is still challenging enough. Our goal was to glean context information and make it available to the application with the lowest latency possible, and indeed, on average the diarization was finished after 0.5 s, as will be described below.

While the online low-latency processing was clearly central to our research, we see the following additional contributions of our work: we present a speaker localization approach based on our recently developed blind beamforming method [22] which achieves in our setup superior accuracy compared to the widely used Generalized cross correlation with phase transform (GCC-PHAT) method. Further we propose to incorporate position and speaker change information via time-variant transition probabilities of a hidden Markov model (HMM) for speaker identification. The resulting joint speaker segmentation and identification achieves good diarization accuracy at low latency. Parts of this work we have presented in earlier publications [11], [23], [24].

Finally, we consider the developed ambient communication system a key contribution of our work, which is built on top of a context management software framework. The task of a context management system is to connect "context sources" to "context consumers" (applications which need or ask for context information), to maintain context information and to merge low-level data to obtain a view of the context at a higher level of abstraction. Following a recent trend in software engineering, where application-based software architectures are replaced by service-oriented designs [25], the context management system proposed here is realized as a service. This increases flexibility and the chance of software reuse, as monolithic applications are done away with and replaced by a set of services which can be composed in various ways to realize different applications. For the realization of the application described in this paper we have employed the service-oriented middleware developed within the Amigo project [26].

The paper is organized as follows. In Section II we will first give an overview of the diarization system, before we describe the individual system components, namely acoustic localization, speaker change detection and identification, as well as face identification, in detail in Section III. In Section IV, our approach to joint segmentation, localization, and identification from streaming audio-visual data will be presented. The key component is a hidden Markov Model with time-variant transition probabilities. Finally, in Section V, we describe the overall software architecture, employing a web service-based context management system. We then demonstrate how the gathered context information is used in an ambient communication system to realize a "follow-me" scenario, where the camera and microphone arrays are automatically steered to the moving user and where the communication follows the user from one room to another. We finish up with some conclusions drawn in Section VI.

II. System Overview

The speaker diarization system integrates components for speaker change detection, speaker identification, speaker localization, and face identification. In Fig. 1, the block diagram of the system is depicted. It is divided in a video subsystem that performs face detection and identification and an audio subsystem that localizes and identifies the speakers. The video system incorporates a single steerable camera, while the audio system contains multiple microphone arrays.

The audio signals captured by the microphones are sampled at 16 kHz and windowed with a block size of 128 samples and a frame advance of 64 samples for the subsequent frame-based audio processing. The video subsystem works also on a frame-by-frame basis; however, here the frame rate depends on the quality of the network, since the employed is network-based. In order to exchange information, both subsystems are connected and synchronized via a shared memory (SHM). Each subsystem uses the information of the other subsystem that is currently stored in the SHM, until it is overwritten by more recent data.

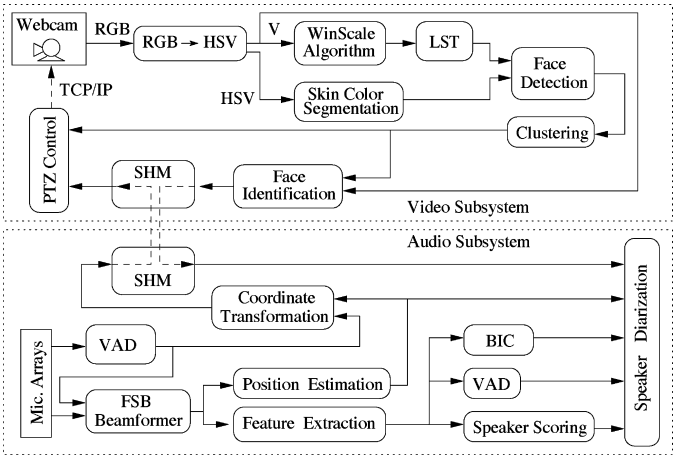

Fig. 1. System overview.

A. Video Subsystem

The purpose of the video subsystem is twofold, as it acts both as a source and a sink of context information. As a source of context information it has to detect and identify the user who is currently in the focus of the camera. As we assume a cooperative user who is facing the camera, this task is achieved by a face detection and identification module. The obtained information is then used for the task of speaker diarization as will be explained later on.

On the other hand, the video system has to focus the camera on the actual user, such that the resulting camera view is optimal for communication. This second task requires context information about the user's location which is based on both the video and audio signals. Position data about the currently active speaker is gathered by the acoustic scene analysis in the audio subsystem and passed through the shared memory to the video subsystem. The module responsible for the camera's pan, tilt and zoom adjustment (*PTZ Control*) compares the information coming from the audio and the video subsystems and fuses them as follows: if both modalities deliver approximately the same information about the user's position, the camera is focused accordingly. In the case that the audio subsystem localizes the user at a different position than indicated by the face detection, the audio localization is favored. This enables the usage of the system by multiple persons during a teleconversation, as the momentarily active speaker is focused even if he is outside the current field of view.

B. Audio Subsystem

The audio subsystem processes the multiple audio streams of the microphone arrays to estimate the identity and position of the user. First of all, a beamformer is employed on each array signal for speech enhancement. As a byproduct of the beamforming step the position of the speaker can be estimated from the filter coefficients, as will be described in more detail in Section III-B. The position estimates are transformed into the coordinate system of the camera and passed to the video subsystem using the shared memory.

The ETSI advanced front-end (AFE) [27] is employed for feature extraction on the output signal of one of the beamformers. It delivers a 13-dimensional Mel-frequency cepstral coefficients (MFCCs) feature vector which is augmented by first- and second-order derivatives. Additionally, a voicedness feature is obtained from an autocorrelation analysis [28]. The resultant feature vector is denoted by $\boldsymbol{x}^{\mathrm{sid}}(k)$, to indicate that it is used for speaker identification in the speaker scoring module. Here, k is the frame index. Further, the feature vector $\boldsymbol{x}^{\mathrm{sid}}(k)$ is passed to the speaker change detection module, which hypothesizes speaker changes based on the Bayesian information criterion (BIC) [29], see Section III-C.

The speaker diarization uses the speaker scores of the speaker identification in combination with the position information and the speaker change hypotheses to estimate the position and identity of the actual speaker. This context information is offered to other applications and services via a web service interface.

Within the audio subsystem two voice activity detections (VADs) are employed. One is used for controlling the adaptation of the beamformer, whose requirements are a low false alarm rate and a moderate missed detection rate. The other is utilized for speaker diarization, which has similar requirements as automatic speech recognition, i.e., a particularly low missed detection rate. For the latter, we employed the voice activity detection of the ETSI AFE, while the former is based on [30].

III. System Components

In the following, we discuss the major building blocks of the audio and video subsystem one-by-one and give an indication of their individual performance.

A. Face Detection and Identification

The video subsystem receives the video stream from the network-based webcam and transforms it from RGB (red, green, blue) to HSV (hue, saturation, value) color space for a skin color segmentation. The regions identified as containing skin color are then examined for faces while the rest of the picture is not considered. At first, the captured frames are gradually scaled down to subframes using the WinScale algorithm from [31]. Subsequently, a local structure transformation (LST) is applied to the subframes [32] and each transformed subframe is scanned for faces by a four-stage detection cascade as proposed by Viola and Jones [33]. Since this face detection approach tends to find a face multiple times a clustering is employed on the found faces to summarize them to single faces.

Next, a region of the grayscale picture around the detected face position is cut out for the following face identification. The classifier itself is based on the well-known Fisherfaces method [34], which uses a principal component analysis (PCA) and a linear discriminant analysis (LDA) to gather the $\mathcal{I}$-dimensional feature vector $\boldsymbol{x}^{\mathrm{vid}}(k)$, where $\mathcal{I}$ is the number of faces, for which trained models exist.

We employ a stochastic approach to the speaker diarization problem, where we model the sequence of feature vectors as realizations of stochastic processes. The class conditional density $p(\boldsymbol{x}^{\mathrm{vid}}(k)|\Omega(k) = i), i \in \{1, \ldots, \mathcal{I}\}$ is assumed to be a Gaussian, whose parameters are obtained from training data. The performance of the face identification is improved by jointly evaluating several observations from the same view angle, which are consecutive in time. To do so, we use the *a*

posteriori probabilities of the last time step as *a priori* probabilities for the current time step. Let $\boldsymbol{x}_{\nu:k}^{\mathrm{vid}} = \boldsymbol{x}^{\mathrm{vid}}(\nu), \ldots, \boldsymbol{x}^{\mathrm{vid}}(k)$ denote the feature vectors of the $(k-\nu+1)$ frames preceding and including the kth frame. Assuming independent and identically distributed (i.i.d.) observations, the posterior probability of class $\Omega(k) = i$ can be recursively computed as

$$P\left(\Omega(k) = i|\boldsymbol{x}_{\nu:k}^{\mathrm{vid}}\right) = \frac{p(\boldsymbol{x}^{\mathrm{vid}}(k)|\Omega(k) = i)P\left(\Omega(k) = i|\boldsymbol{x}_{\nu:k-1}^{\mathrm{vid}}\right)}{\sum_j p(\boldsymbol{x}^{\mathrm{vid}}(k)|\Omega(k) = j)P\left(\Omega(k) = j|\boldsymbol{x}_{\nu:k-1}^{\mathrm{vid}}\right)}. \quad (1)$$

This recursion is started at a frame ν, where a face is detected at a certain position for the first time.

We tested our face identification system on the Yale database [35] to validate its performance. We achieved a minimum error of 4% which is slightly better than the results reported in [34]. The difference is probably due to the fact that we have chosen the parts of the picture for face identification differently than in [34].

B. Acoustic Localization

We consider an array of M microphones. Each discrete-time microphone signal is given by

$$x_i(n) = h_i(n) * s(n) + w_i(n) \quad (2)$$

where $h_i(n)$ denotes the room impulse response, $s(n)$ the desired speaker signal, and $w_i(n)$ additive noise at the ith microphone. We employ a filter-sum beamformer (FSB), whose output signal is given by

$$y(n) = \sum_{i=1}^{M} f_i(-n) * x_i(n). \quad (3)$$

Here, $f_i(n)$ is the impulse response of the beamforming filter in the ith microphone branch and "$*$" denotes convolution. The adaptation of the FSB filter coefficients is based on an eigenvalue decomposition and does not require an estimate of the direction of arrival (DoA) of the desired signal. It blindly adapts to the dominant sound source and can be viewed as a kind of "self-steering" delay-and-sum beamformer; see [22] for details.

A DoA estimate, however, can be obtained as a byproduct of the adaptation as follows. First the cross-correlation as a function of lag λ between any two filter impulse responses is computed as

$$\phi_{ij}(\lambda) = f_i(-\lambda) * f_j(\lambda), \quad (i,j) \in \{1,\ldots,M\}; i \neq j. \quad (4)$$

Since the filters have the ability to model delays which are not an integer multiple of the sampling period T, a continuous-time function $C_{ij}^{\mathrm{FSB}}(\tau)$ can be obtained from $\phi_{ij}(\lambda)$ by interpolation. An estimate of the time delay between the ith and jth microphone signal is then obtained by

$$\tau_{ij} = \underset{\tau}{\operatorname{argmax}} \left\{C_{ij}^{\mathrm{FSB}}(\tau)\right\} \quad (5)$$

which corresponds to the DoA estimate

$$\alpha_{ij} = \arcsin\left(c \cdot T \frac{\tau_{ij}}{d_{ij}}\right) \quad (6)$$

where c is the speed of sound, d_{ij} the distance between ith and jth microphone, and T the sampling period, respectively.

By doing so for any pair of microphones we eventually arrive at the DoA estimate

$$\alpha^{(\mathrm{FSB})} = \frac{2}{M(M-1)} \sum_{i=1}^{M-1} \sum_{j=i+1}^{M} \alpha_{ij}. \quad (7)$$

While with a single linear array only a DoA can be estimated, a position estimate is obtained if L arrays are employed, $L > 1$. Therefore, we calculate beams with directions corresponding to the DoA estimates $\alpha_l^{(\mathrm{FSB})}, l \in \{1,\ldots,L\}$ and starting at the known array positions. Then the speaker position is estimated as the centroid of all beam intersection points (see also Fig. 8 in Section V-B).

The estimate can be somewhat improved if the shape of the cross-correlation function is taken as an indicator of the reliability of the corresponding delay estimate. We define the weighting factor

$$\gamma_{ij} = \frac{\max_\lambda\{|\phi_{ij}(\lambda)|\}}{\sum_{\lambda'} |\phi_{ij}(\lambda')|}. \quad (8)$$

A large value of γ_{ij} indicates a clear maximum which points to a reliable estimate. An improved position estimate can then be obtained by using this reliability information as weights when computing the centroid of the intersection points.

We compare the quality of the localization by the beamformer approach against the well-known GCC-PHAT method [36], which uses the function

$$\phi_{ij}^{\mathrm{GCC}}(\lambda) = \mathrm{IDFT}\left\{\frac{\mathrm{DFT}\{x_i(n)\} \cdot \mathrm{DFT}^*\{x_j(n)\}}{|\mathrm{DFT}\{x_i(n)\} \cdot \mathrm{DFT}^*\{x_j(n)\}|}\right\} \quad (9)$$

to replace $C_{ij}^{\mathrm{FSB}}(\tau)$ in (5), after it has been interpolated to $C_{ij}^{\mathrm{GCC}}(\tau)$.

Both $C_{ij}^{\mathrm{FSB}}(\tau)$ and $C_{ij}^{\mathrm{GCC}}(\tau)$ can be utilized in a global coherence field analysis (GCF) [37]. The global coherence function for a setup of L microphone arrays is given by

$$\mathrm{GCF}(x,y) = \frac{1}{L} \sum_{l=1}^{L} \frac{2}{M_l^2 - M_l} \sum_{i=1}^{M_l-1} \sum_{j=i+1}^{M_l} C_{ij,l}(\tau_{ij,l}(x,y)). \quad (10)$$

In GCF, the room floor is discretized in a finite number of positions (x,y). Each position corresponds to a certain DoA for each microphone array under investigation and thus to a certain delay $\tau_{ij,l}(x,y)$ between any two microphones i and j of the lth array. In (10), the sum is over all microphone pairs of each of the L arrays, where the lth array consists of M_l microphones. A position estimate $(\hat{x},\hat{y})$ is then obtained as that position (x,y) which maximizes (10). Obviously, the number of sampled positions (x,y) has a significant effect on the computational effort and also on the obtainable precision of the localization itself. As a compromise we used a lattice spacing of 0.1 m between two sampled points.

The experiments for acoustic localization utilized the image method from [38] to simulate a room of size 4 $\times$ 4 m and with a height of 3 m and varying room reverberation times T_{60}. Four

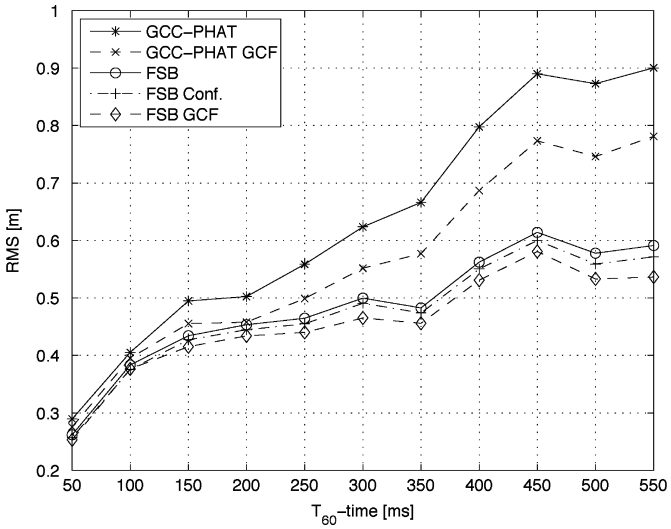

Fig. 2. Comparison of acoustic localization techniques.

microphone arrays, each consisting of two microphones with inter-element distance of 0.05 m, are placed at the center of each wall. We simulated a randomly moving speaker for a 90-s time period, sampling the eight audio streams at 16 kHz. Note that the FFT length for the GCC-PHAT method was set to 2048 and for the FSB to 128.

The experimental results for acoustic localization are depicted in Fig. 2. Obviously, the beamforming approaches FSB, FSB Conf. [FSB with reliability weighting (8)] and FSB GCF (global coherence field analysis with time delays obtained from FSB as input) are superior to the GCC-PHAT techniques (GCC-PHAT, GCC-PHAT GCF). As expected, the computationally more demanding global coherence field analysis (GCC-PHAT GCF, FSB GCF) performs somewhat better than the simpler centroid methods (GCC-PHAT, FSB). This holds for the FSB as well as for the GCC-PHAT results. Furthermore, a small gain for larger T_{60}-times is observable if the centroid method uses the confidence value of (8) as a weighting factor (FSB Conf.).

C. Speaker Change Detection

The speaker change detection is usually the first step to obtain homogeneous segments for speaker identification, clustering, or speech recognition. An overview can be found in [39], with an emphasis on methods employing the BIC.

Speaker change detection can be viewed as a model selection problem and can thus be formulated as a hypothesis test. For a fixed window size of N_w feature vectors, we compare the hypothesis H_0 that all feature vectors $(\boldsymbol{X}_{1:N_w})$ are from one model, against the hypothesis H_1 that the first $N_w/2$ feature vectors $(\boldsymbol{X}_{1:N_w/2})$ are from one model and the rest $(\boldsymbol{X}_{N_w/2+1:N_w})$ from another. According to [3], the difference between the BIC-values of the two hypotheses is given by

$$\Delta\mathrm{BIC} = \frac{N_w}{2}\log(|\boldsymbol{\Sigma}_0|) - \frac{N_w}{4}\log(|\boldsymbol{\Sigma}_1||\boldsymbol{\Sigma}_2|) - \xi'$$

where the covariance matrices $\boldsymbol{\Sigma}_0, \boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$ have to be estimated on the respective feature vector sets $(\boldsymbol{X}_{1:N_w}, \boldsymbol{X}_{1:N_w/2}$ and $\boldsymbol{X}_{N_w/2+1:N_w})$, and ξ' is a constant.

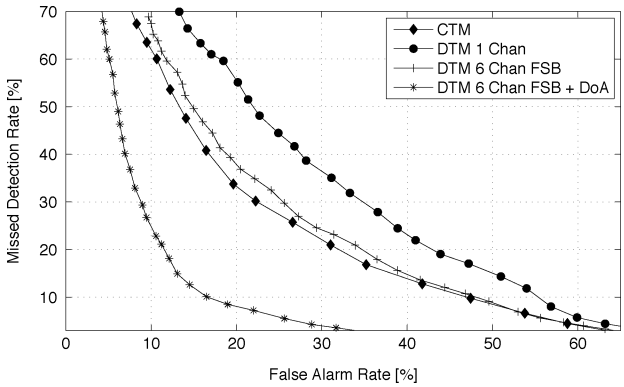

Fig. 3. Speaker change detection results for close (CTM) and distant (DTM) talking microphones.

We prefer to use a fixed window size rather than a growing window as proposed in [40], since this results in a constant latency of $N_w/2$ frames rather than a latency of varying, data-dependent length. However, the window size N_w has to be chosen carefully as a compromise between high segmentation accuracy (large N_w) and low-latency (small N_w). The window is slid over the data, and for each partition k, where k indicates the center frame, a value $\Delta\,\mathrm{BIC}\,(k)$ is computed. Since we are concerned with low-latency processing of streaming data, an iterative refinement of the segmentation cannot be afforded here.

The speaker change detection used in the experiments is based on the approach proposed in [29], where the characteristics of the $\Delta\mathrm{BIC}(k)$ values over time is examined. A peak in the trajectory is called significant and thus indicative of a speaker change if its value exceeds a threshold equal the standard deviation of the $\Delta\mathrm{BIC}(k)$ values times a constant factor.

In Fig. 3, some experimental results on a self-compiled database are given. The database consists of 1.5 hours of spoken texts from a total of five men and five women, who are recorded by distant (DTM) and close talking microphones (CTM) [11]. This database allows a quantitative evaluation of the speaker change detection accuracy and showcases the idea of using location information in speaker change detection. We utilized a self-compiled database, since none of the publicly available databases for speaker change detection experiments fully served our needs, as we needed a database with many speaker changes, short segments and true microphone array recordings.

As expected, the error rates increase if recordings from distant talking microphones (*DTM 1 Chan*) are used instead of the recordings from a close talking microphone (CTM). Since we are interested in localizing and identifying persons that move freely through the house, we have to rely on distant microphone signals. However beamforming for a six-element microphone array with the filter-sum beamformer (*DTM 6 Chan FSB*) improves the performance of the speaker change detection and nearly reaches the performance of the close talking microphone. Further drastic improvements are obtained if DoA information obtained from the microphone array is incorporated into the speaker change detection (*DTM 6 Chan FSB + DoA*). This was done by discarding all speaker changes hypothesized by the BIC analysis, as long as the estimated DoA does not change over

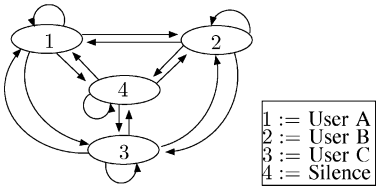

Fig. 4. Hidden Markov model for speaker diarization.

time. The underlying assumption is that different speakers are at different locations and that a speaker does not move abruptly. Thus, a constant DoA indicates that the speaker hasn't changed. This method effectively reduces the number of false alarms and causes only a slight increase of the missed detection rate.

D. Speaker Identification

Our speaker identification system uses Gaussian mixture models (GMMs) with 128 densities, which are adapted by Bayesian learning from gender-dependent background models [41]. These background models are GMMs whose parameters are estimated by the expectation-maximization algorithm from training data. During identification the observed feature vectors are scored against each speaker model and against the fused, i.e., gender-independent, background model. The ratio between the model score and the background model score is employed for the identification task.

IV. Speaker Diarization

A. Hidden Markov Model for Speaker Diarization

Speaker diarization is usually carried out by first segmenting the input data and then assigning speaker identities to the segments. On the contrary, we propose to do the two tasks jointly. To achieve this we employ a HMM, where each of the $\mathcal{I}$ speakers is associated to a hidden state, and an extra state $\mathcal{I}+1$ is introduced to model silence. Fig. 4 shows an example for three users.

A subtlety to be observed is the following. The models for speaker identification have been trained only on data containing speech. Formally, this can be expressed by introducing a voice activity variable V, where $V(k) = 1$ indicates presence of speech, i.e., the HMM state variable $\Omega(k) \in \{1,\ldots,\mathcal{I}\}$ and $V(k) = 0$ indicates $\Omega(k) = \mathcal{I}+1$. Let $p(\boldsymbol{x}^{\mathrm{sid}}(k)|\widetilde{\Omega}(k)), \widetilde{\Omega}(k) \in \{1,\ldots,\mathcal{I}\}$ be the speaker model obtained from training on speech only data. Then the models to be used on input data containing silence are obtained by

$$p(\boldsymbol{x}^{\mathrm{sid}}(k)|\Omega(k)=j) = \begin{cases} p(\boldsymbol{x}^{\mathrm{sid}}(k)|\widetilde{\Omega}(k)=j)\cdot P(V(k)=1) & j\in\{1,\ldots,\mathcal{I}\} \\ p(\boldsymbol{x}^{\mathrm{sid}}(k)|\Omega(k)=j)\cdot P(V(k)=0) & j=\mathcal{I}+1 \end{cases} \tag{11}$$

where the observation probability for a silence frame is set to the average GMM score of the speaker models

$$p(\boldsymbol{x}^{\mathrm{sid}}(k)|\Omega(k)=\mathcal{I}+1) = \frac{1}{\mathcal{I}}\sum_{j=1}^{\mathcal{I}} p(\boldsymbol{x}^{\mathrm{sid}}(k)|\widetilde{\Omega}(k)=j). \tag{12}$$

We define the joint audio-visual "observation probability" of the jth HMM state to be

$$b_j\left(\boldsymbol{x}^{\mathrm{sid}}(k), \boldsymbol{x}^{\mathrm{vid}}_{\nu:k}\right) := p\left(\boldsymbol{x}^{\mathrm{sid}}(k), \boldsymbol{x}^{\mathrm{vid}}_{\nu:k}\right| \Omega(k)=j\big) \tag{13}$$

$$= p(\boldsymbol{x}^{\mathrm{sid}}(k)|\Omega(k)=j) \cdot p\left(\boldsymbol{x}^{\mathrm{vid}}_{\nu:k}|\Omega(k)=j\right) \tag{14}$$

$$= p(\boldsymbol{x}^{\mathrm{sid}}(k)|\Omega(k)=j) \cdot P\left(\Omega(k)=j\,\middle|\,\boldsymbol{x}^{\mathrm{vid}}_{\nu:k}\right)\frac{p\left(\boldsymbol{x}^{\mathrm{vid}}_{\nu:k}\right)}{P(\Omega(k)=j)} \tag{15}$$

with $P(\Omega(k)=j|\boldsymbol{x}^{\mathrm{vid}}_{\nu:k})$ given in (1). Here, we have assumed that $\boldsymbol{x}^{\mathrm{sid}}(k)$ and $\boldsymbol{x}^{\mathrm{vid}}_{\nu:k}$ are independent. In the experiments we also assumed that the *a priori* speaker probabilities $P(\Omega(k)=j)$ are equal for all users. For (13)–(15) we tacitly assumed that the jth class of the face identification refers to the same person as the jth class of the speaker identification task. This is indeed the case, since the camera is focused on the actual speaker using the information from the acoustic position estimate (see also Section II-A). Hence, the speech captured by the microphones and the face seen by the camera must belong to the same person.

It is obvious that the transitions between different states of the HMM should have a higher probability if there is some evidence for a speaker change. We therefore propose to employ time-variant transition probabilities, which are determined by the ΔBIC scores and by changes in the estimated speaker location.

To be specific, we define the binary random variable $c(k)$, which is 1 if a speaker change occurs between the time instances $k-1$ and k, and 0 else. Since peaks in the ΔBIC curve are an indicator for speaker changes we selected the variance of successive ΔBIC values within a time window as one basis for estimating the transition probabilities. To do so, we learned a Gaussian $p(x^{\mathrm{bic}}(k)|c(k)=1)$ and a Gaussian $p(x^{\mathrm{bic}}(k)|c(k)=0)$ beforehand from training data, where x^{bic} is an estimate of the variance of $\Delta\mathrm{BIC}(k)$ determined as follows:

$$\mu^{\mathrm{bic}}(k) = \alpha\mu^{\mathrm{bic}}(k-1) + (1-\alpha)\Delta\mathrm{BIC}(k) \tag{16}$$

$$x^{\mathrm{bic}}(k) = \beta x^{\mathrm{bic}}(k-1) + (1-\beta)[\Delta\mathrm{BIC}(k) - \mu^{\mathrm{bic}}(k)]^2 \tag{17}$$

Here, α and β are smoothing factors. Note that we have employed an exponentially weighted time window extending into the past, over which the variance is estimated to avoid any delay which would be introduced by a centered window.

Similarly, a probabilistic modeling of the position information is established by estimating the variance $x^{\mathrm{pos}}(k)$ of the speaker position from training data just as in (16)–(17), however with $\Delta\mathrm{BIC}(k)$ replaced by the distance between successive position estimates. The position variance is then again modeled as a random variable $x^{\mathrm{pos}}(k)$ with $p(x^{\mathrm{pos}}(k)|c(k)=1)$ and $p(x^{\mathrm{pos}}(k)|c(k)=0)$ being Gaussian densities.

If $x^{\mathrm{bic}}(k)$ and $x^{\mathrm{pos}}(k)$ are assumed to be statistically independent, we get

$$P(c(k)|x^{\mathrm{pos}}(k), x^{\mathrm{bic}}(k)) = \frac{p(x^{\mathrm{pos}}(k), x^{\mathrm{bic}}(k)|c(k))P(c(k))}{p(x^{\mathrm{pos}}(k), x^{\mathrm{bic}}(k))} \quad (18)$$

$$= \frac{p(x^{\mathrm{pos}}(k)|c(k))P(c(k))}{p(x^{\mathrm{pos}}(k))} \frac{p(x^{\mathrm{bic}}(k)|c(k))P(c(k))}{p(x^{\mathrm{bic}}(k))} \cdot \frac{1}{P(c(k))}. \quad (19)$$

Under the assumption of a uniform prior $P(c(k))$ we have

$$P(c(k)|x^{\mathrm{pos}}(k), x^{\mathrm{bic}}(k)) \quad (20)$$

$$= \frac{p(x^{\mathrm{pos}}(k)|c(k))}{\sum_{c'} p(x^{\mathrm{pos}}(k)|c(k) = c')} \cdot \frac{p(x^{\mathrm{bic}}(k)|c(k))}{\sum_{c'} p(x^{\mathrm{bic}}(k)|c(k) = c')} \frac{1}{P(c(k))}. \quad (21)$$

The time-variant transition probabilities between the HMM states are now defined to be

$$a_{ij}(k) = P(\Omega(k) = j|\Omega(k-1) = i) = \frac{\tilde{a}_{ij}(k)}{\sum_j \tilde{a}_{ij}(k)} \quad (22)$$

with

$$\tilde{a}_{ij}(k) = \begin{cases} P(c(k) = 0|x^{\mathrm{pos}}(k), x^{\mathrm{bic}}(k)) & i = j, j \neq \mathcal{I}+1 \\ P(c(k) = 1|x^{\mathrm{pos}}(k), x^{\mathrm{bic}}(k)) & i \neq j, j \neq \mathcal{I}+1 \\ P(c(k) = 0|x^{\mathrm{bic}}(k)) & i = j = \mathcal{I}+1 \\ P(c(k) = 1|x^{\mathrm{bic}}(k)) & i \neq j, j = \mathcal{I}+1 \end{cases}. \quad (23)$$

Since a speaker change can only be indicated through acoustic position estimation, if the user speaks, no position estimates are available for the silence state. Accordingly transition probabilities for transitions to or within the silence state have to be independent from position estimates.

B. Viterbi Decoder

A Viterbi decoder is used to find the single best state sequence $\hat{\Omega}_{1:N} = \hat{\Omega}(1), \ldots, \hat{\Omega}(N)$, given the acoustical and visual observations of length N frames

$$\hat{\Omega}_{1:N} = \arg\max_{\Omega_{1:N}} \left\{ \sum_{k=1}^{N} \left[\log p\left(\boldsymbol{x}^{\mathrm{sid}}(k), \boldsymbol{x}^{\mathrm{vid}}_{\nu;k} | \Omega(k)\right) + \kappa \log P(\Omega(k)|\Omega(k-1)) \right] \right\}. \quad (24)$$

The constant κ is introduced to weigh the observation probabilities against the transition probabilities. By this, excessive state switching can be suppressed. The Viterbi decoder is implemented with a partial traceback to reduce the latency entailed by the decoder.

C. Experiments With Audio Data

In this first set of experiments, we are going to investigate the usefulness of employing time-variant transition probabilities and a Viterbi decoder for speaker diarization [23]. To this end we leave out the video subsystem and concentrate on the performance of the audio subsystem. The database used is the same as the one for the speaker change detection experiments described above. However, we grouped the audio files according to the average segment length.

TABLE I
DIARIZATION ERROR RATES OF DIFFERENT SETUPS

Method \ Duration	Diarization Error Rate [%] $< 2s$	$3-4s$	$> 4s$	Avg.
ΔBIC	28.76	13.91	7.94	12.98
Viterbi (Fixed)	25.53	10.05	5.72	9.66
Viterbi (Pos)	21.66	9.32	5.69	8.95
Viterbi (BIC)	24.03	9.48	5.35	9.08
Viterbi (Pos,BIC)	22.80	6.80	4.27	7.05
Ground-truth change pts	11.09	4.05	2.46	4.00

Table I gives an overview of the simulation results for speaker diarization. The performance of the diarization methods is measured with the diarization error rate (DER), which is the percentage of incorrectly labeled frames [3]. Since we know the optimal segmentation points ("Ground-truth change pts") we are able to determine the lower bound of the diarization error rate which is given by the non-perfect performance of the speaker identification. The baseline, denoted by ΔBIC, is the standard approach for segmentation via BIC and subsequent identification of the speaker.

The application of a Viterbi decoder with fixed transition probabilities ("Viterbi (Fixed)") improves the diarization results compared to the ΔBIC method, although it does not incorporate any speaker change information. Next we utilized only one knowledge source for estimating the transition probabilities and in both cases ("Viterbi (Pos)" and "Viterbi (BIC)") the diarization performance is increased. The best performance is achieved when position and speaker change information are the basis for estimating the transition probabilities ("Viterbi (Pos, BIC)").

D. Experiments With Audio-Visual Data

The video subsystem is a highly dynamic component, as the orientation and focus of the camera is controlled by the face detection and acoustic speaker localization.

Let us first visualize the time variant behavior of the posterior probabilities given by (1) by means of an example. The upper half of Fig. 5 shows the evolution of the *a posteriori* probabilities over time, while the lower half displays the location estimate of the acoustic scene analysis in Cartesian coordinates (x, y).

At time instant 3.9 s, a speaker change happens, which can be clearly seen by the rapid change of the estimated x-coordinate. Since the camera still observes the face of the previous speaker its focus remains at the old position for a short time until it starts focusing the new speaker. This idle period is introduced to avoid uncontrolled movements of the camera in case of short-term acoustic interferences, e.g., slamming doors. If a new speaker is indicated outside the view of the camera for some minimum number of position estimates, the camera's view turns to the new speaker even if faces are detected in the current viewpoint. In the example of Fig. 5, the camera needs the time from approximately 4 s until 8 s to focus the new speaker. During

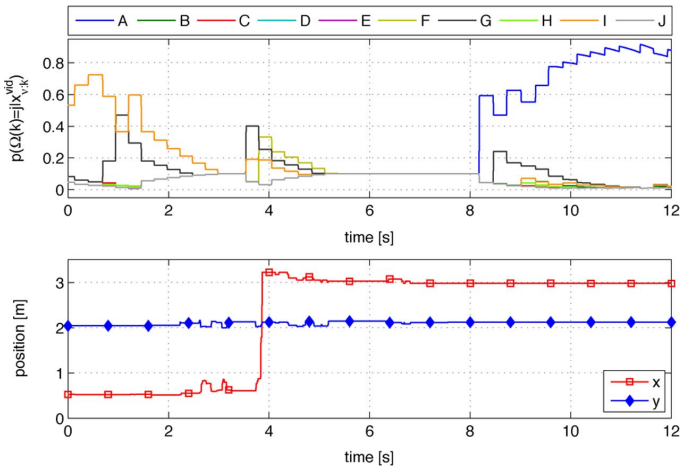

Fig. 5. Example of behavior of *a posteriori* probabilities of faces (upper figure) and acoustic positioning estimates (lower figure) over time.

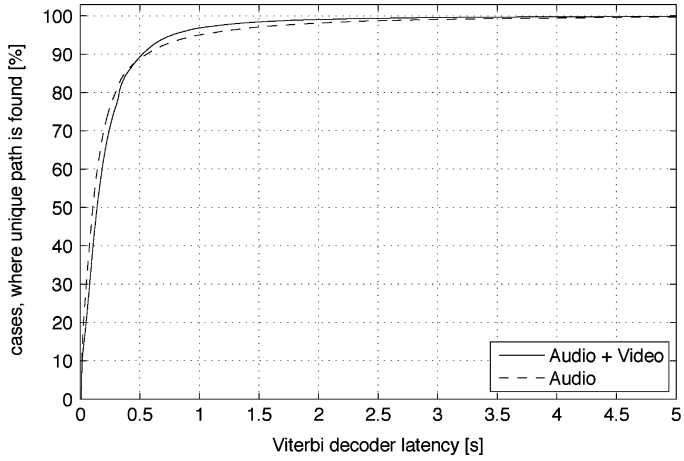

Fig. 6. Percentage of cases, where the unique path is found versus decoding latency.

TABLE II
EXPERIMENTAL RESULTS FOR AUDIO-VISUAL IDENTIFICATION

Case	User	Faces [%] obs.	Faces [%] corr.	DER [%] audio	DER [%] fusion	time [min:sec]
Examples single user	A	83.55	83.99	5.13	2.96	3:07
	B	72.51	83.97	6.22	4.67	7:43
	C	94.18	74.60	16.54	11.65	3:18
	D	94.27	100.00	24.88	1.13	2:57
	E	93.70	19.51	6.58	14.41	2:47
	F	56.16	90.30	7.91	1.38	6:27
Examples multiple users	A & D	75.99	82.76	24.56	7.81	3:14
	A & B	88.56	82.84	33.79	5.22	3:36
	C & D	89.03	86.48	15.45	8.23	7:38
	D & E	75.65	74.17	14.79	12.67	6:09
	A & F	52.90	89.84	34.25	9.78	3:31
	B & D	60.49	41.68	23.50	15.07	5:47
Average single		84.53	84.79	7.46	3.72	61:18
Average multiple		76.66	74.08	23.11	11.81	59:24
Average all		80.46	79.49	15.16	7.70	120:42

this period of time, the *a posteriori* probabilities are equal for all users. After that, the probability of one of the users sharply rises, while the others gradually decrease.

To quantify the improvements gained by integrating the observations of the video subsystem into the speaker diarization process we conducted several experiments. First, we used the system with single users who were at fixed positions in the room for most of the time. We considered this to be typical for a telecommunication scenario. Next we examined multi-user cases, which could be representative of a conferencing scenario, where on one side of the telecommunication system several persons are present. The total number of users, for which the system was trained, was $\mathcal{I} = 10$.

In the third column of Table II, the percentage of time is given, in which faces have been detected. The fourth column contains the identification accuracy for the detected faces. As expected, in the multi-user case faces are less frequently found due to the time the camera needs for focusing the new speaker. The results for speaker diarization can be found in the fifth and sixth column, where the diarization error rates (DER) for audio-based diarization ("audio") and audio-video-based diarization ("fusion") are compared. In the last column, the duration of the experiments are given. We selected some experiments from the total amount of 2 hours of data, which are representative for the system performance. The average results for the full database are given in the last three rows.

As can be inferred from Table II, the system performs reasonably well for the single user case. Since the user's movement is limited a very low DER is archived for audio-video based diarization. For user "D" the benefit of the additional visual information is most obvious. If the face identification accuracy is low, e.g., in case of user "E," video information can be counterproductive and even increase the diarization error rate. If multiple users are present the DER is higher compared to the single user case. Although less reliable information from the video subsystem is available it still results in a considerable improvement of the error rate. On average the combination of audio and visual information reduces the diarization error rate from 15.16% to 7.70% compared to the audio-only results.

E. Latency Considerations

The application presented in Section V has tight latency requirements. In the following, we give some considerations related to latency and its impact on the diarization process.

At first some latency is caused by the connection between the soundcard and the software. In the optimal case (real-time operating system) it equals the frame size of the audio processing unit, which is 128/16 000 s $=$ 8 ms in our system. The subsequent audio processing by the beamformer and the ETSI AFE is virtually free of latency and instantly delivers a feature vector for each input block.

The ΔBIC values are computed for a window of $N_w = 80$ frames, resulting in a latency of $N_w/2 \cdot 8$ ms $= 320$ ms. Only after this period a decision on a speaker change can be made. The computation of the time-variant transition probabilities does not entail any additional latency.

The Viterbi decoder used in the speaker diarization has a variable latency. In Fig. 6, the percentage of cases, where the unique path was found by the traceback procedure is depicted versus the corresponding latency of the Viterbi decoder. It can be observed that in 90% of the cases the traceback finds a unique path after 0.5 s. Compared with the audio only case, the speaker diarization with audio and video information needs on average somewhat less time until a unique path is found.

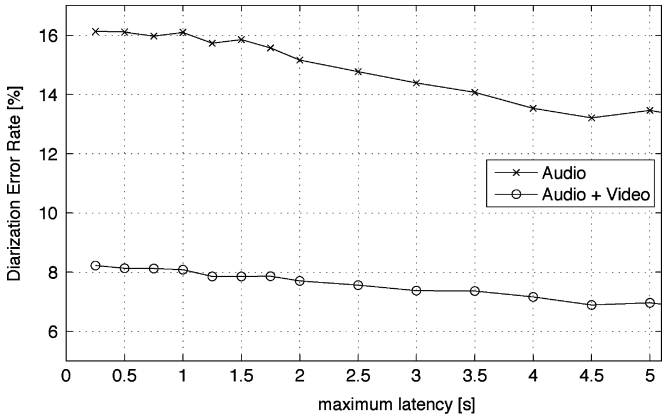

Fig. 7. Diarization error rate versus maximum latency of the Viterbi decoder after which a decision is enforced.

In theory, the delay introduced by the Viterbi decoder can be arbitrarily large. Therefore, a maximum delay has to be introduced, after which an output is enforced. In Fig. 7, we show the dependence of the diarization error rate on the maximum delay. As expected the error rate decreases with an increasing time limit for the traceback. We decided that a maximum delay of 2 s could be an appropriate tradeoff between latency and speaker diarization accuracy.

The average time until the unique path in the trellis is found was experimentally found to be 246.2 ms for the fusion of audio and video information and 262.2 ms for the audio only case, while the median delay is 104 ms and 136 ms, respectively. Summing up all latencies, the average latency of the system for speaker diarization is 566.2 ms in the case video information is available and 582.2 ms else, while the main contribution of 320 ms is due to BIC. Disregarding the speaker change information based on BIC would reduce the latency to about a quarter of a second; however, the price to be paid would be an increased diarization error rate; see Table I.

V. Integration Into an Ambient Communication System

The speaker diarization is a valuable context source for applications in smart environments as it provides information about user location and identity. However, this information is useless if it is not offered and distributed to applications. For this reason, a middleware is required that provides some basic functionality for context distribution and service interaction. First, we will give a brief description of the middleware and afterwards we will introduce an example service that uses the speaker diarization as a context source.

A. Middleware Technologies

The open source middleware of the Amigo project is targeted at smart environments, with a focus on the networked home environment [26]. Since the configuration of services, applications and devices can be quite different from one home to another, the middleware has to be flexible and cater for different infrastructures and usage preferences [42]. This is achieved by the service-oriented architecture (SOA) of Amigo, that is based on interacting services.

Moreover, to account for interoperability among devices of different vendors and to integrate established middleware technologies the Amigo middleware is based on ontologies and web services [42], [43] as inspired by the ideas of the semantic web [44]. All ontologies are formulated in the Web Ontology Language (OWL) [45], which uses a resource description framework/extensible markup language (RDF/XML) [46] notation.

Services providing a web service interface [47] are described either in a semantic way by the *Amigo-S* language [48] or purely syntactically by the *Web Services Description Language* (WSDL) [45]. *Amigo-S* is an extended version of the *Web Ontology Language for Web Services* (OWLS) [50] that is enlarged by *Quality of Service* (QoS) and context-awareness aspects.

One of the key components of the Amigo middleware is the *Context Management Service* (CMS), which is responsible for bringing together services providing context information and the context consumers, e.g., applications or devices [51]. A service that intends to use the Amigo middleware to publish context information has to implement a web service interface for interaction with other services, and it has to announce its capabilities to the CMS. Applications searching for context sources define their requirements in a request to the CMS and subsequently receive a list of suitable context sources. Then the applications can either directly ask for a certain piece of context information or register for notifications at the context source. In both cases, the queries to the services are formulated in the *SPARQL Protocol and RDF Query Language* (SPARQL) [52] and the answers are placed in RDF/XML format. The Amigo middleware is running on an *Open Services Gateway Initiative* (OSGI) platform [53]. The whole Amigo software suite is available as open source software [54].

B. Ambient Communication

Ambient communication follows the vision of AmI systems in that devices are integrated in the environment. The telecommunication is supposed to be carried out in hands-free mode and to be continued without the user's interaction even if he moves from one room to another [55]. On top of that, the communication may exist for an extended period of time and at varying levels of intensity not asking for full attention all of the time. Ambient Communication thus supports the feeling of staying or living together despite a potentially large physical distance [13].

In Fig. 8, our experimental setup for the ambient communication system is shown, where the output devices are a display and a loudspeaker and the input devices being a steerable camera and three microphone arrays. Array_1 is a T-shaped four-element microphone array which allows for the estimation of an azimuthal angle α_1 and a tilt angle β.

The figure is also meant to illustrate the acoustic localization described in Section III-B. Together with the other two-element arrays Array_2 and Arrray_3 a total of three DoAs $\alpha_i, i = 1, 2, 3$ are estimated, from which the position $(\hat{x}, \hat{y})$ can be estimated as the centroid of the intersection points s_{13}, s_{23} and s_{12}.

Fig. 9 shows the major building blocks we used to set up an ambient communication system. It is split into the two parts "Speech & Video I/O" and "Middleware," which are connected via an UDP-based (user datagram protocol) interprocess communication (IPC). In addition to providing basic

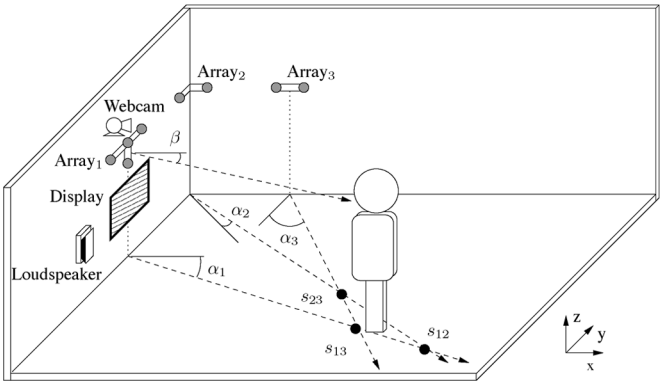

Fig. 8. Experimental setup illustrating the placement of audio-visual devices.

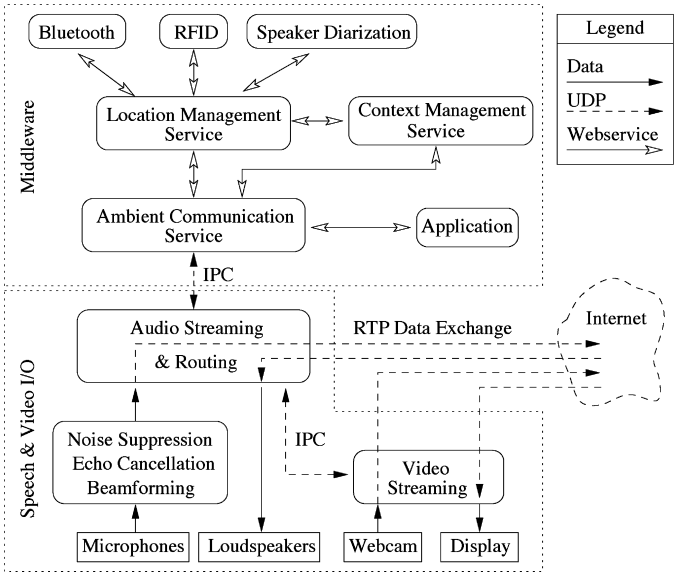

Fig. 9. Overview of ambient communication system.

speech input/output functionality, the speech and video I/O part does some speech enhancement by conducting acoustic beamforming and multi-channel echo cancellation, followed by a residual echo and noise suppression postfilter. On top of the signal processing, a module for audio streaming and routing is implemented, which is controlled by the ambient communication service via an IPC connection. It is responsible for the real-time streaming session handling to distant sites and audio stream routing within the house. The streaming utilizes a 16-kHz wideband Speex coder [56] and the real-time transport protocol (RTP) [57]. Video data from the camera is compressed with a Theora coder [58] and exchanged with a second RTP stream parallel to the audio stream. Thereby the audio streaming module controls the video streaming module via an IPC.

The "smartness" of the ambient communication system is realized within the middleware, since all position data and other communication session relevant information are available there. Instead of connecting to all available context sources that provide position information, the ambient communication service registers to the location management service. This service permanently searches for new context sources providing position data via the context management service, registers to them and combines the different information to a consistent database of user locations. Position data may originate from bluetooth devices, radio frequency identification (RFID) location systems or the speaker diarization described in this paper and thus may vary in precision, reliability, and temporal resolution.

The ambient communication service is not only a consumer of context information, it is also a context providing service. Data about active connections, users available for communication and the hardware equipment of the rooms represent contextual information for applications that intend to use the communication service. Therefore, the ambient communication service integrates a second web service interface for context exchange besides the web service interface for session control.

Applications interested in setting up audio connections search for ambient communication services via the context management system and register to the context delivering web service interface. If a suitable service is found, a connection between two sites is set up via a web service call to the communication service. The communication itself is internally bound to the users, which implies that the application does not have to take care of position changes of the user. If a user moves from one room to another, the ambient communication service is informed about the context change by the location management service. This triggers the streaming modules to redirect the audio-video data to the corresponding room.

While Bluetooth or RFID localization mostly have a precision at the room level, acoustic localization is more precise and able to locate the user within a room. Thus, in a typical setup the wireless RFID localization system tracks room changes of the user to redirect the audio-visual data to the entered room, while the speaker diarization locates the speaker within the room.

A public demonstration of the ambient communication system was given in February 2008 during the Amigo Project Open Day in Eindhoven (The Netherlands). Visitors were invited to use the system in communication scenarios with sites in France and Germany, where different innovative aspects could be tested. For instance the site in Germany showed the audio-visual camera focusing and the site in France presented proximity-based services [7]. In Eindhoven, the follow-me functionality between rooms had been realized, such that it could be directly experienced by the visitors.

VI. Conclusion

While speaker diarization is traditionally concerned with attributing temporal regions to specific speakers given single-channel audio recordings, we extended the task both with respect to the goal and the given input data, to better suit applications in smart environments. The goal was extended to retrieve also speaker location information and to deliver the diarization result with a latency as low as possible. Further, the input was assumed to consist of streaming audio-visual data from multiple microphone arrays and a steerable camera. From these acoustic position and face identification results are obtained as additional knowledge sources for the diarization task.

In this paper, we have shown how to combine these additional information sources with speaker change detection probabilities obtained from a BIC-based hypothesis test and acoustic

speaker identification information in a joint segmentation, localization, and identification system. The key component is an HMM with time-variant state transition probabilities which are derived from speaker and position change probabilities, and observation probabilities which combine acoustic and visual user identification evidence. The experiments have shown that the position information significantly reduces the diarization error rate (DER), and that the visual face identification leads to a reduction of the DER by roughly a factor of two.

Special emphasis has been placed on reducing the latency after which the diarization result is available: low latency has been achieved by carefully designing the BIC-based speaker change detection and by decoding the HMM with a Viterbi decoder incorporating a partial traceback and enforcing a decision after a time-out. An average latency of about half a second is obtained. The latency can be further reduced by about a factor of two, if BIC-based speaker change detection is removed, however at the cost of an increased diarization error rate.

The diarization system processes the sensor data online and provides the user's location and identity as context information to an Ambient Intelligence application. As an example, we realized an ambient communication system, i.e., a system for audio-visual telecommunication, where the user can move freely within his home without carrying a device and the communication follows him using the most appropriate I/O devices according to the provided context information. The system was realized using the open source service-oriented middleware developed within the Amigo project. It was shown during the Amigo Project Open day to the general public.

REFERENCES

[1] E. Aarts and S. Marzano, *The New Everyday: Views on Ambient Intelligence*. Rotterdam, The Netherlands: 010 Uitgeverij, 2004.

[2] "Information Society Technologies Advisory Group Reports," 2003 [Online]. Available: http://cordis.europa.eu/fp7/ict/istag/reports_en.html

[3] S. Tranter and D. Reynolds, "An overview of automatic speaker diarization systems," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 5, pp. 1557–1565, Sep. 2006.

[4] S. Meignier, D. Moraru, C. Fredouille, J. Bonastre, and L. Besacier, "Step-by-step and integrated approaches in broadcast news speaker diarization," *Comput. Speech Lang.*, vol. 20, no. 2–3, pp. 303–330, Jul. 2006.

[5] C. Wooters and M. Huijbregts, "The ICSI RT07s speaker diarization system," in *Proc. Multimodal Technol. Percept. Humans: Int. Eval. Workshops CLEAR 2007 and RT 2007*, Baltimore, MD, May 8–11, 2007, pp. 509–519, 2008 Revised Selected Papers.

[6] X. Zhu, C. Barras, L. Lamel, and J. Gauvain, "Multi-stage speaker diarization for conference and lecture meetings," in *Lecture Notes in Computer Science: Machine Learning for Multimodal Interaction, no. 4625*, 2008, pp. 533–542.

[7] S. Borkowski and G. Privat, "Spatial interaction in ambient communication," in *Proc. Int. Conf. Enactive Interfaces (ENACTIVE'07)*, Grenoble, France, Nov. 2007.

[8] H. K. Ekenel, M. Fischer, Q. Jin, and R. Stiefelhagen, "Multi-modal person identification in a smart environment," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognition (CVPR'07)*, Minneapolis, MN, Jun. 2007, pp. 1–8.

[9] A. Salah, R. Morros, J. Luque, C. Segura, J. Hernando, O. Ambekar, B. Schouten, and E. Pauwels, "Multimodal identification and localization of users in a smart environment," *J. Multimodal User Interfaces*, vol. 2, no. 2, pp. 75–91, Sep. 2008.

[10] A. Noulas and B. J. A. Krose, "On-line multi-modal speaker diarization," in *Proc. Int. Conf. Multimodal Interfaces (ICMI'07)*, New York, Apr. 2007, pp. 350–357.

[11] J. Schmalenstroeer and R. Haeb-Umbach, "Online speaker change detection by combining bic with microphone array beamforming," in *Proc. Conf. Int. Speech Commun. Assoc. (Interspeech'06)*, Pittsburgh, PA, Sep. 2006.

[12] J. Pardo, X. Anguera, and C. Wooters, "Speaker diarization for multiple-distant-microphone meetings using several sources of information," *IEEE Trans. Comput.*, vol. 56, no. 9, pp. 1212–1224, Sep. 2007.

[13] S. Borkowski, T. Flury, A. Gerodolle, and G. Privat, "Ambient communication and context-aware presence management," *Commun. Comput. Inf. Sci.: Constructing Ambient Intell.*, vol. 11, pp. 391–396, 2008.

[14] A. Härmä, "Ambient telephony: Scenarios and research challenges," in *Proc. Conf. Int. Speech Commun. Assoc. (Interspeech'07)*, Antwerp, Belgium, Aug. 2007.

[15] "Computers in the human interaction loop," CHIL Consortium Jan. 2004 [Online]. Available: http://chil.server.de/

[16] "Augmented multi-party interaction," AMI Consortium Jan. 2004 [Online]. Available: http://www.amiproject.org/

[17] A. Temko, R. Malkin, C. Ziegler, D. Macho, C. Nadeu, and M. Omologo, "Acoustic event detection and classification in smart-room environments: Evaluation of CHIL project systems," *J. Tecnologia del Habla*, vol. 4, pp. 1–6, Nov. 2006.

[18] X. Anguera, C. Wooters, and J. Hernando, "Acoustic beamforming for speaker diarization of meetings," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 7, pp. 2011–2022, Sep. 2007.

[19] "Detection and identification of rare audiovisual cues," DIRAC Consortium Jan. 2006 [Online]. Available: http://www.diracproject.org/

[20] T. Kühnapfel, T. Tan, S. Venkatesh, and E. Lehmann, "Calibration of audio-video sensors for multi-modal event indexing," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP'07)*, Honolulu, HI, Apr. 2007, pp. 741–744.

[21] D. Lo, R. A. Goubran, R. M. Dansereau, G. Thompson, and D. Schulz, "Robust joint audio-video localization in video conferencing using reliability information," *IEEE Trans. Instrum. Meas.*, vol. 53, no. 4, pp. 1132–1139, Aug. 2004.

[22] E. Warsitz and R. Haeb-Umbach, "Acoustic filter-and-sum beamforming by adaptive principal component analysis," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP'05)*, Philadelphia, PA, Mar. 2005, pp. 797–800.

[23] J. Schmalenstroeer and R. Haeb-Umbach, "Joint speaker segmentation, localization and identification for streaming audio," in *Proc. Conf. Int. Speech Commun. Assoc. (Interspeech' 07)*, Antwerp, Belgium, Aug. 2007.

[24] J. Schmalenstroeer, M. Kelling, V. Leutnant, and R. Haeb-Umbach, "Fusing audio and video information for online speaker diarization," in *Proc. Conf. Int. Speech Commun. Assoc. (Interspeech'09)*, Brighton, U.K., Sep. 2009.

[25] M. J. Carey, "SOA what?," *IEEE Trans. Comput.* , vol. 41, no. 3, pp. 92–94, Mar. 2008.

[26] "Ambient intelligence for the networked home environment," 2006 [Online]. Available: http://www.hitechprojects. com/euprojects/amigo

[27] *Speech Processing, Transmission and Quality Aspects (Stq); Distributed Speech Recognition; Advanced Frontend Feature Extaction Algorithm; Compression Algorithms*, ETSI. (2007) Es 202 212 v1.1.5 [Online]. Available: http://www.etsi.org

[28] B. Wildermoth and K. Paliwal, "Use of voicing and pitch information for speaker recognition," in *Proc. IEEE Conf. Speech Sci. Technol. (SST'00)*, Canberra, Australia, Dec. 2000, pp. 324–328.

[29] P. Delacourt and C. J. Wellekens, "Distbic: A speaker-based segmentation for audio data indexing," *Speech Commun.*, vol. 32, no. 1–2, pp. 111–126, Sep. 2000.

[30] J. Ramirez and J. Segura, "Efficient voice activity detection algorithms using long-term speech information," *Speech Commun.*, vol. 42, no. 3–4, pp. 271–287, Apr. 2004.

[31] C. Kim, S. Seong, J. Lee, and L. Kim, "Winscale: An image-scaling algorithm using an area pixel model," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 13, no. 6, pp. 549–553, Jun. 2003.

[32] B. Froeba and C. Kueblbeck, "Face tracking by means of continous detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognition Workshop (CVPRW'04)*, Washington, DC, Mar. 2004, pp. 65–71.

[33] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognition (CVPR'01)*, Kauai, HI, Dec. 2001, pp. 511–518.

[34] P. Belhumeur, J. Hespanha, and D. Kriegman, "Eigenfaces vs. fisherfaces: Recognition using class specific linear projection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, no. 7, pp. 711–720, Jul. 1997.

[35] "The Yale Face Database," Yale University. New Haven, CT, Apr. 2009 [Online]. Available: http://cvc.yale.edu/projects/yalefaces/yalefaces.html

[36] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-24, no. 4, pp. 320–327, Aug. 1976.
[37] M. Omologo, P. Svaizer, A. Brutti, and L. Cristoforetti, "Speaker localization in CHIL lectures: Evaluation criteria and results," *Lecture Notes in Computer Science, no. 3869*, pp. 476–487, 2006.
[38] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Amer.*, vol. 65, no. 4, pp. 943–950, Apr. 1979.
[39] M. Kotti, V. Moschou, and C. Kotropoulos, "Speaker segmentation and clustering," *Signal Process.*, vol. 88, no. 5, pp. 1091–1124, May 2007.
[40] S. S. Chen and P. S. Gopalakrishnan, "Speaker, environment and channel change detection and clustering via the bayesian information criterion," in *Proc. DARPA Broadcast News Transcription and Understanding Workshop*, Lansdowne, VA, Feb. 1998.
[41] D. Reynolds, T. Quatieri, and R. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital Signal Process.*, vol. 10, no. 1–3, pp. 19–41, Jan. 2000.
[42] D. Sacchetti, Y. Bromberg, N. Georgantas, V. Issarny, J. Parra, and R. Poortinga, "The amigo interoperable middleware for the networked home environment," in *Proc. Middleware*, Grenoble, France, Dec. 2005.
[43] N. Georgantas, S. B. Mokhtar, Y. Bromberg, V. Issarny, J. Kalaoja, J. Kantarovich, A. Gerodolle, and R. Mevissen, "The amigo service architecture for the open networked home environment," in *Proc. Working IEEE/IFIP Conf. Software Architecture (WICSA'05)*, Pittsburgh, PA, Nov. 2005.
[44] T. Berners-Lee, J. Hendler, and O. Lassila, "The semantic web," *Sci. Amer. Mag.*, pp. 1–4, May 2001.
[45] D. L. McGuinness and F. van Harmelen, "Web ontologie language," World Wide Web Consortium Feb. 2004 [Online]. Available: http://www.w3.org/TR/owl-features/
[46] D. Beckett *et al.*, "Resource description framework," World Wide Web Consortium, Jan. 2008 [Online]. Available: http://www.w3.org/RDF/
[47] WWW. (2002) Web Services. World Wide Web Consortium. [Online]. Available: http://www.w3.org/2002/ws/
[48] S. Mokhtar, A. Kaul, N. Georgantas, and V. Issarny, "Efficient semantic service discovery in pervasive computing environments," *Lecture Notes in Computer Science, no. 4290*, pp. 240–259, 2007.
[49] R. Chinnici *et al.*, "Web services description language," World Wide Web Consortium. Jun. 2007 [Online]. Available: http://www.w3.org/TR/wsdl20/
[50] DAML. (2006) Web Ontology Language for Web Services. DARPA Agent Markup Language. [Online]. Available: http://www.daml.org/services/owl-s/
[51] F. Ramparany, R. Poortinga, M. Stikic, J. Schmalenstroeer, and T. Prante, "An open context information management infrastructure," in *Proc. IET Int. Conf. Intell. Environments (IE'07)*, Ulm, Germany, Sep. 2007.
[52] E. Prud'hommeaux and A. Seaborne, "SPARQL protocol and RDF query language," World Wide Web Consortium, Jan. 2008 [Online]. Available: http://www.w3.org/TR/rdf-sparql-query/
[53] Open Service Gateway Initiative, Jan. 2008 [Online]. Available: http://www.osgi.org
[54] F. Ramparany *et al.*, "Amigo software repository," Amigo Consortium, 2009 [Online]. Available: http://amigo.gforge.inria.fr
[55] J. Schmalenstroeer, V. Leutnant, and R. Haeb-Umbach, "Amigo context management service with applications in ambient communication scenarios," *Commun. Comput. Inf. Sci.: Construct. Ambient Intell.*, vol. 11, pp. 397–402, 2008.
[56] Audio Codec, 2008 [Online]. Available: http://www.speex.org
[57] H. Schulzrinne *et al.*, "RTP: A Transport Protocol for Real-Time Applications. Internet Engineering Task Force," Jul. 2003 [Online]. Available: http://tools.ietf.org/html/rfc3550
[58] "Theora Codec," 2008 [Online]. Available: http://www.Theora.org

Joerg Schmalenstroeer received the Dipl.-Ing. and Dr.-Ing. degree in electrical engineering from the University of Paderborn, Paderborn, Germany, in 2004 and 2010, respectively.

Since 2004, he has been a Research Staff Member with the Department of Communications Engineering, University of Paderborn. His research interests are in acoustic scene analysis and statistical speech signal processing.

Reinhold Haeb-Umbach (M'89) received the Dipl.-Ing. and Dr.-Ing. degrees in electrical engineering from RWTH Aachen University, Aachen, Germany, in 1983 and 1988, respectively.

From 1988 to 1989, he was a Postdoctoral Fellow at the IBM Almaden Research Center, San Jose, CA, conducting research on coding and signal processing for recording channels. From 1990 to 2001, he was with Philips Research working on various aspects of automatic speech recognition, such as acoustic modeling, efficient search strategies, and mapping of algorithms on low-resource hardware. Since 2001, he has been a Professor in communications engineering at the University of Paderborn, Paderborn, Germany. His main research interests are in statistical speech signal processing and recognition and in signal processing for communications. He has published more than 100 papers in peer-reviewed journals and conferences.