Audio-visual Data Processing for Ambient Communication

Joerg Schmalenstroeer[1], Volker Leutnant[2], Reinhold Haeb-Umbach[3]

University of Paderborn, Department of Communications Engineering, Germany
[1]schmalen@nt.uni-paderborn.de,[2]leutnant@nt.uni-paderborn.de,
[3]haeb@nt.uni-paderborn.de

Abstract. In this paper we present our system for acoustic scene analysis and ambient communication. The acoustic scene analysis delivers information about the user's location which is utilized in ambient communication such that audio-visual data are captured and rendered by the most appropriate I/O-device, which allows the user to move freely from one room to another during a teleconversation. The system employs a steerable camera, controlled jointly by acoustic speaker localization and face detection. The ambient communication system is implemented on top of a context management system which maintains context information provided by context sources and consumed by applications.

1 Introduction

Ambient communication as a future trend of ambient telephony [1] formulates the vision of a user-oriented, service based infrastructure for audio and video communication [2]. Thus, it follows the paradigm of Ambient Intelligence (AmI) that claims the key elements of an "intelligent" system to be embedded, context-aware, personalized, adaptive and anticipatory [3]. Hence it overcomes the limitations of a hardware-oriented telephony application or device by hiding the hardware from the user in the walls and at the same time retaining and extending its original functionality.

The aforementioned elements of AmI can only be realized if a sufficient amount of reliable context information is available. This constraint asks for two tasks to be solved by an intelligent system. First of all sensors, devices and applications have to be integrated in a network, utilizing a common middleware for communication and interaction. Second, the inherent knowledge of the information sources has to be prepared such that machines can understand and process it. A widely accepted approach for this task is the use of an ontology, which in principle constitutes a joint knowledge base by definition of terms and their relationships.

Regarding ambient communication scenarios the acoustic signals recorded by the microphones are interesting context sources as they provide information about users and events. Obviously localization and identification by audio signals assumes that the user is speaking, however during a communication this should

be fulfilled. Our acoustic scene analysis localizes and identifies active speakers and thus generates information about: "Who speaks, When and Where?". This context information is used by a camera to focus the active speaker and it is also provided to other applications.

In the next section we briefly describe the usage scenario and the used hardware. Section 3 gives an overview about the system building blocks for audio and video processing. After presenting the middleware in section 4, section 5 explains our system for ambient communication, and we finish with some conclusions.

2 Usage scenario

We envisage a networked home environment as the typical environment where ambient telephony is to be used. It is characterised by a multiplicity of hardware components stemming from the domains consumer electronics, household appliances, personal computing and telecommunication, all more or less connected via networks. Especially, if we focus on the audio-visual equipment, we find a large variety of hardware configuration to be installed in the home. This may range from single or no microphones per room to rooms equipped with distributed microphone arrays, loudspeakers and cameras.

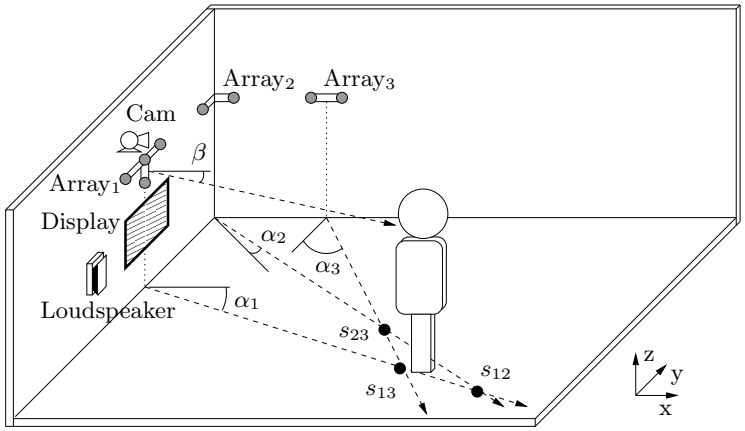

Fig. 1. Ambient communication setup

Our further explanations are based on a room with a high level of equipment as depicted in Fig. 1. For audio signal processing three microphone arrays, namely one T-shaped and two linear arrays, are used. The T-shaped one is mounted at the wall between a display and a pan-tilt-zoom camera. It is assumed that the user looks in the direction of the camera, and thus in the direction of the array, while having an audio-visual communication with a distant person. Together with the two other arrays the speaker can be located which is intern used to focus the camera on the user.

3 System overview

The system for audio-visual data processing is divided in two parts working in parallel, which are synchronized and connected via a shared memory (SHM) approach, see Fig. 2. In the video subsystem the webcam stream is processed on a frame-by-frame basis, where the frame rate may vary because of changing network quality. The audio subsystem works at a constant sampling rate of 16 kHz and a block length of 10 ms. Information gathered by one of the subsystems is stored in the shared memory and used by the other until it is overwritten.

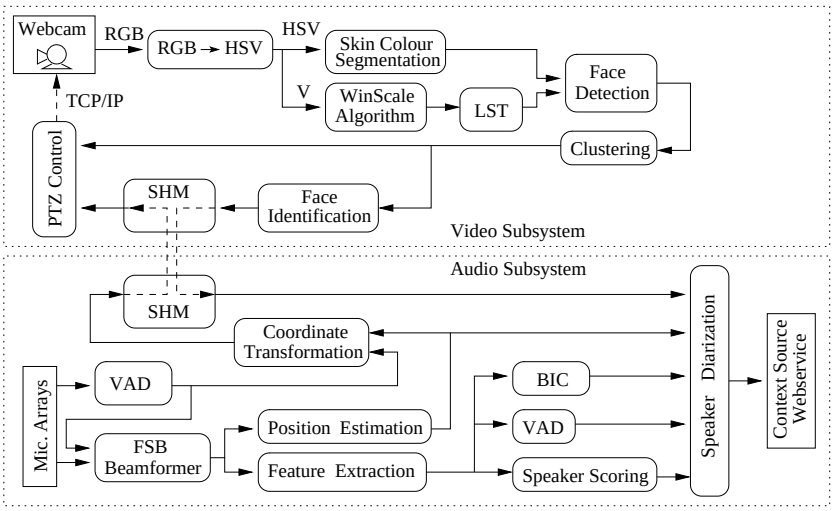

Fig. 2. Speaker localization and camera control

3.1 Video subsystem

The frames of the video stream are converted from RGB to HSV to ease the skin color segmentation and in parallel retrieve the grey scale version (V component) of the frames. We perform a histogram look-up to find the regions of skin color, which simultaneously reduces the computational demand and the false alarm rates of the face detector by constraining the areas of the frame to be examined for faces. We employ a face detector that is optimized to find faces at a size of 19×19 pixels. Thus we have to scale down the original frame to subframes with different resolutions to find faces at larger sizes. Here we employ the WinScale algorithm from [5]. Each subframe is processed with a local structure transformation (LST) as proposed in [6] and subsequently scanned for faces by a 4-stage detection cascade as suggested by Viola and Jones in [7].

The face detection method tends to detect a face multiple times in marginally varying positions and sizes, thus a Leader-Follower clustering is employed to

merge the results. According to the information from the face detection we cut out the parts of the greyscale picture at the face positions and scale them to a size of 60×60 pixels. Further the well-known Fisherfaces approach [8] is applied to identify the persons. In a first step we use a principal component analysis (PCA) matrix that was determined on training data to reduce the feature vector size from 3600 to 200. In the second step we further reduce the dimension to the number of trained users minus one, by applying a LDA matrix from a linear discriminant analysis (LDA) that was also estimated on the training data. A single Gaussian is estimated for each user to model him in a probabilistically manner. Consecutive observations of faces in the same look direction are tracked by interpreting the posterior likelihoods of the last timestep as a priori likelihoods of the current timestep. The current posteriors are stored in the shared memory.

3.2 Audio subsystem

The audio subsystem uses the spatially distributed microphones for localization and identification of speakers (cf. Fig. 2). First of all we use a beamformer for speech enhancement to reduce the detrimental effects of reverberation and noise. We employed a filter-sum beamformer (FSB) [9] which performs a principal component analysis on each microphone array signal and thus blindly adapts to the strongest sound source. The correlation of the FSB filter coefficients enables an estimation of the Direction-of-Arrival (DoA) for each array and jointly a localization of the user, if multiple distributed arrays are available. In our setup the DoA information of each array is transformed in corresponding azimuth angles $[\alpha_1, \alpha_2, \alpha_3]$ while the T-shaped array is also able to provide a tilt angle estimate β (cf. Fig. 1). Next we calculate the intersection points $[s_{13}, s_{23}, s_{12}]$ of the direction estimates and retrieve the speaker position estimate as their centroid.

Speaker identification requires a segmentation of the audio stream in homogeneous parts. Since the applications in mind asks for online data processing with short latencies, multi-stage batch procedures or iterative methods as normally proposed for speaker diarization [10] are not applicable. Our approach uses a Hidden Markov Model (HMM) where each state corresponds to a certain user. A partial traceback is implemented to enable joint speaker segmentation, identification and localization at low latency [11]. In contrast to other methods, e.g. [12], we estimate a time variant transition matrix from speaker change hypotheses. Information about possible speaker changes are retrieved from the variance of the speaker localization and the variance of the Bayesian Information Criterion (BIC) [13].

We use the ETSI advanced feature extraction front-end [14] on the beamformer output signal to to obtain a 39-dimensional feature vector. The vector is extended to 42 dimensions by adding a voicedness feature and its first and second order derivatives. The speaker scoring calculates the likelihoods from the feature vectors, based on the Gaussian mixture models (GMM) of the users. Further we interpret the posteriors of the face identification as a priori knowledge for the speaker diarization. It follows that the product of the GMM likelihoods and the

posteriors of the face identification are the state observation probabilities of the HMM. The partial traceback of the speaker diarization module estimates the single best state sequence given the acoustical and visual observations and then hands over the information to the context source. This context source can be used by any application or device via its webservice interface.

3.3 Camera control

We employed a pan-tilt-zoom camera for visual communication and also for identifying users. The camera orientation and depth view is controlled by incorporating visual as well as acoustical position information. Localized users that are not within the camera view are automatically focused so that the currently active speaker gets into the camera view after just a short delay.

4 Middleware and context management

The middleware represents the backbone of a networked home environment. Its ability to provide context information and to integrate different services is of utmost importance to realize a perceived level of "intelligence". Our system builds upon the open source middleware that was developed during the Amigo project [15]. It uses webservice technologies from the semantic web and comes along with basic services for context management, aggregation and distribution [16]. In Fig. 3 the architecture of the Amigo context management system is depicted. The context broker (CB) is the central unit for registering and searching context sources, whereas context sources are defined as any element that delivers a kind of information that may be interesting for the system.

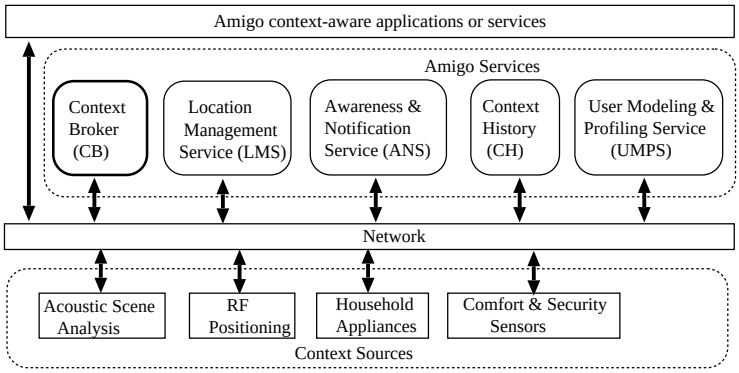

Fig. 3. Amigo context management system

Applications or services search via the context broker for suitable context sources and contact them by their standardized webservice interfaces. Either a

direct request for information is forwarded to the context source or the application registers at the context source for notifications in case of new context information. In both cases a SPARQL query is formulated [17] and the answer is given in RDF/XML description format [18].

A key context information is the user's location, which is handled by the Amigo middleware within the location management service (LMS). This service continuously searches for context sources providing location information, e.g. the acoustic speaker diarization or a RFID positioning system. All context information is aggregated by the LMS and delivered as new contextual information to other applications.

5 Webservice audio interface

The webservice based audio interface connects the audio processing part for communication with the context-aware applications using the Amigo middleware. We coined this assembly of building blocks *Seamless Audio Interface* (SAInt) to outline one of the key features of the system. SAInt realizes a follow-me functionality for audio communications such that the user can freely change rooms while the communication follows him seamlessly.

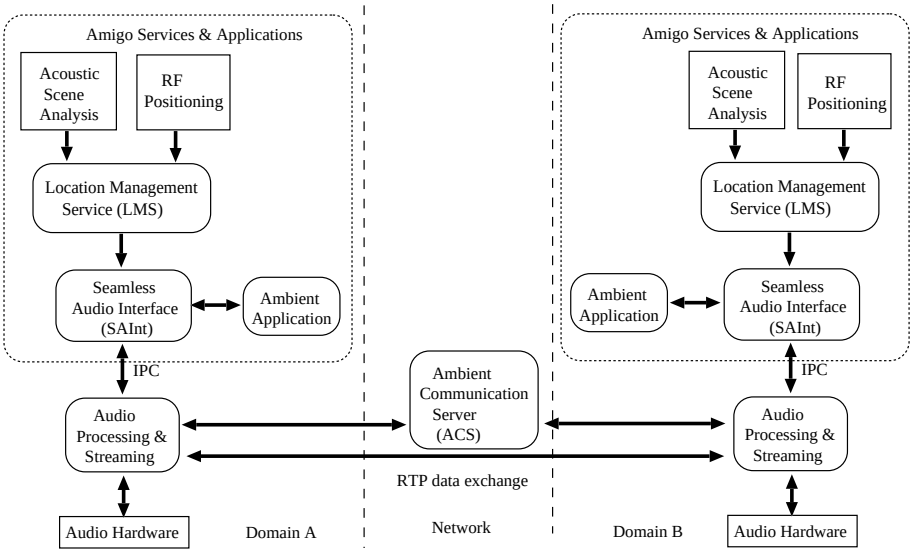

Fig. 4. SAInT: Seamless audio interface

In Fig. 4 the building blocks of SAInt are depicted. The audio processing and streaming block receives the acoustical signals from the sound capturing hardware and first of all performs an echo cancellation and noise suppression for

signal enhancement. The streaming itself is initiated, controlled and terminated by the applications or by contextual information. Therefore the signal processing block is asynchronously connected via an interprocess communication (IPC) with the SAInt middleware service.

SAInt obtains information about user locations directly from the LMS and offers a webservice interface for applications. In parallel SAInt acts as a context source, publishing information about the rooms equipped with audio hardware, about ongoing connections and about users available for communication. Thus applications can get an overview about the hardware and the users in range of it by registering to all SAInt services in the connected home.

An application asks for an audio or audio-visual connection by instructing the SAInt service to connect two persons. SAInt uses the LMS to look up the location of the persons and sets up the connection. If a person moves from one room to another, the change of context information triggers a redirect of the audio streaming, while the application using SAInt does not have to take care about it. Thus a communication is internally bound to a user and follows him on his way through the house.

The audio streams are compressed with an 16 kHz Speex wideband audio coder and the video data is compressed with the Theora coder, both including a packet-loss concealment. We use the real-time transport protocol (RTP) for interchanging the audio and video data between two SAInt instances. External connections to other houses are initiated via a central server that is called *ambient communication server* (ACS). It enables firewall and network address translation (NAT) traversal as well as session initialization and handovers.

6 Discussion

In this paper we have presented our system for ambient communication and acoustic scene analysis. Both tasks are closely related, as they are based on the same acoustical signals. We have shown how context information about the user's location is obtained from analyzing the data captured by microphone arrays and a steerable camera. This location information is intern utilized to control the camera and to select the most appropriate I/O-device while the user is moving freely in the home doing a teleconversation with a remote partner. We have further described an open middleware which connects context sources to context consumers and which thus enables services to take context-aware decisions.

References

1. A. Härmä, "Ambient Telephony: scenarios and research challenges", Proceedings Interspeech 2007, Antwerpen, 2007
2. S. Borkowski, T. Flury, A. Gerodolle, G. Privat, "Ambient Communication and Context-Aware Presence Management", Communications in Computer and Information Science, Vol. 11, pp. 391-396, Springer LNCS, Germany, Berlin, 2008

3. S. Marzano, E. Aarts, “The New Everyday - Views on Ambient Intelligence”, Koninklijke Philips Electronics N.V., 010 Publishers, The Netherlands, Rotterdam, 2004
4. N. Georgantas, S. B. Mokhtar, Y. Bromberg, V. Issarny, J. Kalaoja, J. Kantarovich, A. Gerodolle, R. Mevissen, “The Amigo Service Architecture for the Open Networked Home Environment”, Proceedings 5th Working IEEE/IFIP Conference on Software Architecture (WICSA), 2005
5. C. Kim, S. Seong, J. Lee, L. Kim, “WinScale: An Image-Scaling Algorithm Using an Area Pixel Model”, IEEE Transactions on Curcuits and Systems for Video Technology, Vol. 13, No.6, pp. 549-553, 2003
6. B. Froeba, C. Kueblbeck, “Face tracking by Means of Continuous Detection”, Proceedings IEEE Computer Society Conference on Computer Vision and Pattern Recognition, USA, Washington, DC., 2004
7. P. Viola, M. Jones, “Rapid Object Detection using a Boosted Cascade of Simple Features”, Proceedings IEEE Conference on Computer Vision and Pattern Recognition, Hawaii, Kauai, 2001
8. P. Belhumeur, J. Hespanha, D. Kriegman “Eigenfaces vs. Fisherfaces: Recognition Using Class Specific Linear Projection”, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 19, No. 7, pp. 711-720, Jul. 1997
9. E. Warsitz, R. Haeb-Umbach “Acoustic Filter-and-Sum Beamforming by Adaptive Principal Component Analysis”, Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing, Philadelphia, USA, 2005
10. S. Tranter, D. Reynolds , “An overview of Automatic Speaker Diarization Systems”, IEEE Transactions on Audio, Speech and Language Processing, Vol. 14, No. 5, pp. 1557-1565, Sep. 2006
11. J. Schmalenstroeer, R. Haeb-Umbach “Joint Speaker Segmentation, Localization and Identification for Streaming Audio”, Proceedings Interspeech 2007, Belgium, Antwerp, 2007
12. S. Meignier et al., “Step-by-Step and Integrated Approaches in Broadcast News Speaker Diarization”, Computer Speech Language, Vol. 20, Issues 2-3, pp. 303-330, Sept. 2005
13. M. Nishida, T. Kawahara, “Speaker Model Selection Based on the Bayesian Information Criterion Applied to Unsupervised Speaker Indexing”, IEEE Transactions on Speech and Audio Processing, Vol. 13, No. 4, pp. 583-592, July 2005
14. ETSI ES 202 050 V1.1.3 , ”ETSI Standard Speech Processing, Transmission and Quality Aspects (STQ); Distributed speech recognition; Advanced front-end feature extraction algorithm; Compression algorithms”, Nov. 2003
15. Amigo Project website: http://www.amigo-project.org, 2007
16. F. Ramparany, R. Poortinga, M. Stikic, J. Schmalenstroeer, T. Prante, “An open Context Information Management Infrastructure - the IST-Amigo Project”, Proceedings Conference on Intelligent Environments, Ulm, Germany, 2007
17. SPARQL Protocol and RDF Query Language: http://www.w3.org/TR/rdf-sparql-query/
18. Resource Description Format (RDF) Specifications: http://www.w3.org/RDF/