Hindawi Publishing Corporation
Advances in Multimedia
Volume 2013, Article ID 175745, 21 pages
http://dx.doi.org/10.1155/2013/175745

Research Article

Real-Time Audio-Visual Analysis for Multiperson Videoconferencing

Petr Motlicek,[1] Stefan Duffner,[1,2] Danil Korchagin,[1] Hervé Bourlard,[1] Carl Scheffler,[1] Jean-Marc Odobez,[1] Giovanni Del Galdo,[3] Markus Kallinger,[3] and Oliver Thiergart[3]

[1] *Idiap Research Institute, 1920 Martigny, Switzerland*
[2] *Université de Lyon, CNRS, INSA-Lyon, LIRIS, UMR5205, 69621 Lyon, France*
[3] *Fraunhofer IIS, 91058 Erlangen, Germany*

Correspondence should be addressed to Stefan Duffner; stefan.duffner@liris.cnrs.fr

Received 28 February 2013; Accepted 21 June 2013

Academic Editor: Alexander Loui

We describe the design of a system consisting of several state-of-the-art real-time audio and video processing components enabling multimodal stream manipulation (e.g., automatic online editing for multiparty videoconferencing applications) in open, unconstrained environments. The underlying algorithms are designed to allow multiple people to enter, interact, and leave the observable scene with no constraints. They comprise continuous localisation of audio objects and its application for spatial audio object coding, detection, and tracking of faces, estimation of head poses and visual focus of attention, detection and localisation of verbal and paralinguistic events, and the association and fusion of these different events. Combined all together, they represent multimodal streams with audio objects and semantic video objects and provide semantic information for stream manipulation systems (like a virtual director). Various experiments have been performed to evaluate the performance of the system. The obtained results demonstrate the effectiveness of the proposed design, the various algorithms, and the benefit of fusing different modalities in this scenario.

1. Introduction

Together Anywhere, Together Anytime (THETA2) project aims at understanding how technology can help to nurture family-to-family relationships to overcome distance and time barriers. This is something the current technology does not address well. Modern media and communications are designed for individuals, as phones, computers, and electronic devices tend to be user centric and provide individual experiences.

Technological goal of TA2 is to build a system enabling natural remote interaction by exploiting sets of individual state-of-the-art "low-level-processing" audio-visual algorithms combined on a higher level. This paper focuses on the description and evaluation of these algorithms and their combination to be eventually used in conjunction with higher-level stream manipulation and interpretation systems, for example, an orchestrated videoconferencing system [1] that automatically selects relevant portions of the data (i.e., using a so-called virtual director). The aim of the proposed system is to separate semantic objects in the low-level signals (like voices, faces) to be able to determine their number and location, and, finally, determine, for instance, who speaks and when. The underlying algorithms comprise continuous localisation of audio objects and its application for spatial audio object coding [2], detection, and tracking of faces, estimation of head poses and visual focus of attention, detection and localisation of verbal and paralinguistic events, and the association and fusion of these different events, which are performed on a per room basis. To quantitatively evaluate the individual algorithms as well as the whole real-time/low delay system, experiments have been carried out on two datasets containing high-definition audio and video data recorded in an unconstrained videoconferencing-like environment.

1.1. Related Work. There is a comprehensive literature on algorithms for multiple face detection and tracking, speaker localisation and diarisation, multimodal fusion techniques, and tracking systems. Most of these existing systems are designed for rather constrained environments, like meeting rooms [3], can only work offline (on prerecorded data), or they use a different technical setup (e.g., collocated sensors).

Most existing work focuses predominantly on a single modality (audio or video). For multiple face tracking, many approaches have been presented in the literature and they mainly deal with improving the overall tracking performance by proposing new features or new multicue fusion mechanisms, and results are demonstrated mostly on short sequences or on videos containing only two persons. Particle filters have proven to be an effective and efficient approach for visual object tracking. For instance, one such algorithm for multitarget tracking has been proposed by Khan et al. [4] and is based on reversible-jump Markov chain Monte Carlo (RJ-MCMC) sampling. But to be effective, it requires appropriate global scene likelihood models involving a fixed number of observations (independent from the number of objects) and these are difficult to build in multiface tracking applications.

On the audio analysis side, there are diarisation systems that identify the speech segments corresponding to each speaker ("who spoke when?") and estimate the number of speakers. Conventional speaker diarisation systems [5] use an ergodic Hidden Markov Model (HMM), where the speakers are represented with different HMM states. Good results were achieved by the systems using combination of mel-frequency cepstral coefficients (MFCCs) and time difference of arrival (TDOA) features with arrays composed of a different number of microphones, while the performance of the TDOA features applied separately was poor [6]. TDOA features can be used without prior knowledge of geometry of the microphone array. If the geometry of the microphone array is known in advance, TDOA features can be replaced by the speaker locations, which can be used alone [7], or as complementary features to conventional MFCCs. Typically, speaker localisation can either be done in the audio modality, video modality, or both. The first one implies using a microphone array, while the second one is based on motion detection or person detection. Multimodal localisation allows results to be less affected by noise and reverberation in the audio modality, although it increases significantly the computational complexity.

Finally, the fusion of audio and video cues can be performed at different levels, based on the type of input information available. It can be done at sensor level, feature level, score level, rank level, or decision level. The first two levels can be considered as preclassification, while the others can be considered as postclassification [8]. The feature-level multimodal approach is usually represented by transforming the data in such a way that a correlation between the audio and a specific location in the video is found [9]. In our work, the score-level fusion is used and is based on a technique relying on information derived from spatially separated sensors [10]. Other score-level multimodal techniques rely on the estimation of the mutual information between the average acoustic energy and the pixel value [11], probability densities estimation [12], or a trained joint probability density function [13].

1.2. Challenges and Motivation. The examined TA2 scenario presents several scientific and implementation-related challenges: audio-visual streams recorded at high resolution (i.e., audio channels captured using a microphone array sampled at 48 kHz allowing to represent any kind of acoustic event without perceptual quality loss; video streaming captured with a high-definition camera) and semantic information need to be computed in real time with low delay from spatially separated sensors within a room (as opposed to other systems, such as [14], relying on collocated sensors). Furthermore, the considered environment is open and rather unconstrained. Video processing algorithms hence must take into account a varying number of persons whose positions are not predefined in the room. In audio, any type of generated acoustic event (e.g., overlapping speech, music, distortions due to the room reverberation captured by distant microphones, or background noise) can appear. This poses real challenges for the audio processing components, especially together with an open dictionary as a natural choice towards the automatic recognition of unconstrained speech. Finally, the association and fusion of extracted acoustic and visual events is not a trivial task, because at each time instants there might be some events that are more reliable than others. The combined model has to be able to estimate a confidence of the different modalities, weight them accordingly, and reliably associate them to the detected persons.

The proposed audio-visual system builds on existing state-of-the-art individual audio and video preprocessing blocks which have been developed over a long time using the author's know-how at their institutes. Nevertheless, this paper describes an integration and extension of these individual blocks to eventually perform real-time analysis of complex audio-visual signals/events recorded within high resolution and with distributed sensors. To our knowledge, such a system does exist neither in a commercial sphere nor in research domain.

In the following, we will first briefly present the overall architecture of the system (Section 2). In Section 3, we will describe the intelligent audio capturing. Section 4 outlines the individual algorithms used for semantic information extraction. Section 5 describes evaluation experiments performed on individual blocks as well as on the whole system. We will also briefly analyse the computational costs of the whole system. Section 6 summarises the achieved results and concludes the paper.

2. Architecture

The proposed system processes the audio and video inputs from spatially separated sensors (see Figure 1), located within a room. By placing the sensors at their individually optimal locations (video input is placed further for better scene coverage, while audio inputs are placed closer to participants to allow better intelligibility and localisation), we clearly

obtain a better performance of audio object separation and low-level semantic information.

The system architecture can be grouped into four parts (see Figure 2). The main components of the system are an audio communication engine (ACE, Section 3), a long-term multiple face tracking and person identification (parts of video cue detection engine (VCDE), Section 4.1), head pose and visual focus of attention estimation (parts of VCDE, Section 4.2), visual speaker and speech detection from head motion (part of VCDE, Section 4.3), audio spatial localisation (part of audio cue detection engine (ACDE), Section 4.4), voice activity detection and keyword spotting (parts of ACDE, Section 4.5), and multimodal calibration, association, and fusion (unified cue detection engine (UCDE), Section 4.6). The output of the system consists of audio objects, semantic video objects, and semantic events and states.

3. Intelligent Audio Capture

The intelligent audio capture aims at identifying and extracting the sound sources from microphone recordings and transforming them into individual audio objects. The object-based representation of a recorded sound scene offers great flexibility in terms of sound enhancement, transmission, and reproduction. The main parts of the intelligent audio capturing are depicted in Figure 2 (ACE block) and discussed in detail in the following sections. The system is based on a parametric representation of the recorded spatial sound using the directional audio coding (DirAC) framework [15]. The parametric representation enables an efficient and robust localisation and extraction of the sound sources in a room, which can then be transformed into an object based representation such as MPEG Spatial Audio Object Coding (SAOC) [2].

3.1. Parametric Spatial Sound Representation. The intelligent audio capturing is based on a sound field model which is especially suitable for speech recordings in a reverberant environment. Let us consider a sound field in the short-time frequency domain where the sound pressure $S(k,n)$ with time index n and frequency index k in the recording location is composed of a superposition of *direct sound* and *diffuse sound*, that is,

$$S(k,n) = S_{\text{dir}}(k,n) + S_{\text{diff}}(k,n). \tag{1}$$

The direct sound $S_{\text{dir}}(k,n)$ (corresponding for instance to speech, propagating directly from the speaker to the microphones) equals to a single monochromatic plane wave with mean power $P_{\text{dir}}(k,n) = E\{|S_{\text{dir}}(k,n)|^2\}$ and direction of arrival (DOA) $\varphi(k,n)$. In contrast, the diffuse sound field $S_{\text{diff}}(k,n)$ (corresponding e.g., to the late reverberation) is assumed to be spatially isotropic, meaning that the sound arrives with equal strength from all directions, and spatially homogeneous, meaning that its mean power $P_{\text{diff}}(k,n)$ does not vary with different positions. Such a diffuse field can be modelled, for example, by summing an infinite number of monochromatic plane waves with equal magnitudes, random phases, and uniformly distributed propagation directions.

In the following, $S_{\text{dir}}(k,n)$ and $S_{\text{diff}}(k,n)$ are assumed to be uncorrelated. Therefore, the total sound power is

$$P(k,n) = E\left\{|S(k,n)|^2\right\} = P_{\text{dir}}(k,n) + P_{\text{diff}}(k,n). \tag{2}$$

The power ratio between the direct sound and diffuse sound is expressed by the signal-to-diffuse Ratio (SDR) $\Gamma(k,n)$, that is,

$$\Gamma(k,n) = \frac{P_{\text{dir}}(k,n)}{P_{\text{diff}}(k,n)}. \tag{3}$$

The recorded spatial sound is described via a parametric representation in terms of $S(k,n)$, $\varphi(k,n)$, and the so-called *diffuseness* $\Psi(k,n)$ representing and alternative expression of the SDR $\Gamma(k,n)$, that is,

$$\Psi(k,n) = \frac{1}{1+\Gamma(k,n)}. \tag{4}$$

The diffuseness becomes zero when only the direct sound is present, one when the sound field is purely diffuse and 0.5 when both fields possess equal power. When the diffuseness is known, the power of the direct sound can be determined from the total sound power using (2), (3), and (4), that is,

$$P_{\text{dir}}(k,n) = (1-\Psi(k,n))\,P(k,n). \tag{5}$$

As explained in the following sections, the DOA $\varphi(k,n)$ and diffuseness $\Psi(k,n)$ can be estimated using a B-format microphone or a microphone array [2, 15].

Clearly, the sound field model in (1) requires that only one sound source is active per time-frequency bin (k,n) together with the diffuse sound. This model holds reasonably well for speech applications even in double talk situations when using a filter bank with proper time-frequency resolution for transforming the microphone signals into the short-time frequency domain [16].

3.2. Continuous Localisation System. The ACE block scheme in Figure 2 depicts the main parts of the sound source localisation system which are explained more in detail in the following sections. Inputs to the system are the signals of a microphone array being transformed into the time-frequency domain using a filter bank. More precisely, we consider a 1024-point short-time fourier transform (STFT) with 50% overlap at a sampling frequency of $f_s = 44.1$ kHz, resulting in a frame size of approximately $T = 11.6$ ms. The transformed microphone signals are fed to the parameter estimation block where the DOA $\varphi(k,n)$ and diffuseness $\Psi(k,n)$ of the sound field are determined. Based on the parametric representation, the long-term spatial power density (LT-SPD) is computed representing a power-weighted long-term histogram of the DOA estimates corresponding to the directional sound. Finally, a clustering algorithm is applied to the LT-SPD providing the number $N(n)$ of sound sources and their angular positions $\theta_{1\ldots N}(n)$.

(1) *Parameter Estimation.* The spatial parameters $\varphi(k,n)$ and $\Psi(k,n)$ are estimated based on the active sound intensity

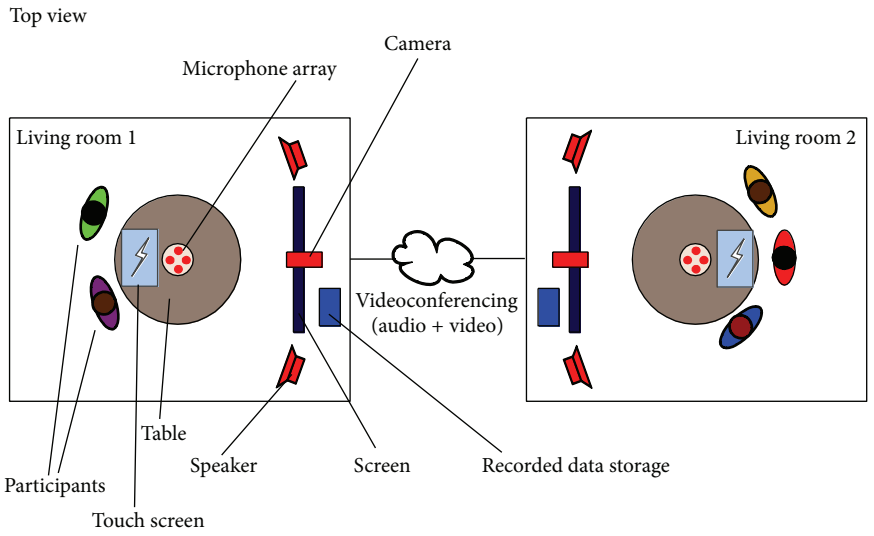


Figure 1: TA2 setup, view from top [36]. The audio and video sensors are spatially separated within a room: the microphone array is located above the table next to participants, while the camera is collocated with the wall screen for teleconferencing.

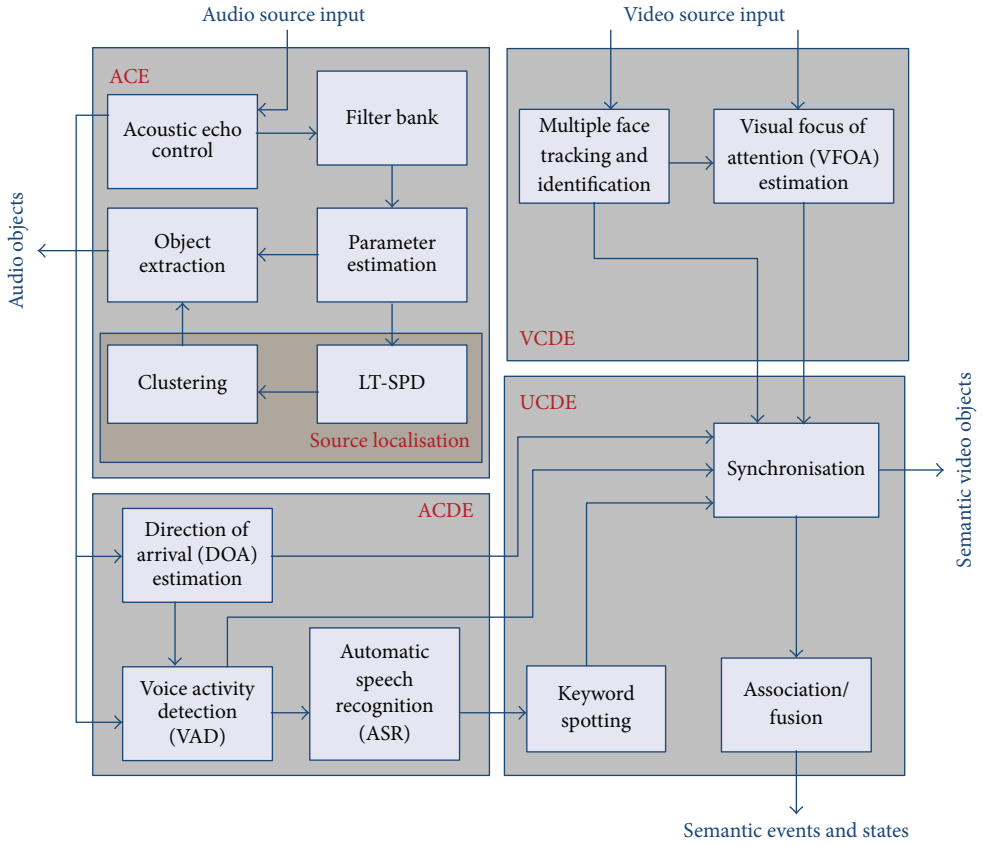


Figure 2: Block diagram of the intelligent audio capturing and semantic information extraction modules. The components are grouped into four parts: audio communication engine (ACE), audio cue detection engine (ACDE), video cue detection engine (VCDE), and unified cue detection engine (UCDE).

as explained in [2, 15]. We employ a planar array of four omnidirectional microphones arranged on the corners of a square with diagonal d. Let $S_i(k,n)$ with $i \in [1,4]$ be one of the four microphone signals in the short-time frequency domain. The components of the active sound intensity vector $\mathbf{I}_a(k,n) = [I_x(k,n)\ \ I_y(k,n)]^T$ describing the net flow of energy in the array center are determined by

$$\begin{aligned} I_x(k,n) &= \mathrm{Re}\left\{W^*(k,n)\ \ X_x(k,n)\right\}, \\ I_y(k,n) &= \mathrm{Re}\left\{W^*(k,n)\ \ X_y(k,n)\right\}, \end{aligned} \tag{6}$$

where $W(k,n) = (1/4)\sum_{i=1}^{4} S_i(k,n)$ is the approximate sound pressure in the array center, with $(\cdot)^*$ denoting complex conjugate, and

$$\begin{aligned} X_x(k,n) &= K\left(S_1(k,n) - S_3(k,n)\right), \\ X_y(k,n) &= K\left(S_2(k,n) - S_4(k,n)\right) \end{aligned} \tag{7}$$

is the approximate particle velocity component along the (x, y) axis of the Cartesian coordinate system, and K is a frequency-dependent complex normalisation factor [2]. The direction of $\mathbf{I}_a(k,n)$ represents the estimated DOA $\widetilde{\varphi}(k,n)$, that is,

$$\frac{\mathbf{I}_a(k,n)}{\left|\mathbf{I}_a(k,n)\right|} = \begin{bmatrix} \cos\widetilde{\varphi}(k,n) \\ \sin\widetilde{\varphi}(k,n) \end{bmatrix}. \tag{8}$$

This estimator provides accurate results for the true DOA $\varphi(k,n)$ of the direct sound for high SDRs $\Gamma(k,n)$. The variance of $\widetilde{\varphi}(k,n)$ increases for lower SDRs, that is, when the sound field becomes more diffuse. In purely diffuse sound fields, $\widetilde{\varphi}(k,n)$ is approximately uniformly distributed within 2π. The behaviour of $\widetilde{\varphi}(k,n)$ as well as of the direction of $\mathbf{I}_a(k,n)$ is further exploited for estimating the diffuseness of the sound. In fact, the diffuseness $\Psi(k,n)$ can be determined via the coefficient-of-variation (CV) of $\mathbf{I}_a(k,n)$ defined as

$$\widetilde{\Psi}(k,n) = \sqrt{1 - \frac{\left|\left\langle \mathbf{I}_a(k,n)\right\rangle_n\right|}{\left\langle\left|\mathbf{I}_a(k,n)\right|\right\rangle_n}}, \tag{9}$$

where $\langle\rangle_n$ denotes temporal averaging. In purely diffuse sound fields, the numerator becomes close to zero leading to unity diffuseness. When only a single plane wave is present, arriving from a fixed direction, the numerator and denominator are equal leading to zero diffuseness. As shown in [17], this estimator represents a close approximation of the definition in (4).

(2) *LT-SPD*. The sound source localisation is based on a power-weighted histogram of the direct sound DOAs similarly to [18]. To obtain this histogram, let us first compute the LT-SPD for different directions $\varphi' \in [-\pi, \pi]$ as

$$\Lambda\left(\varphi', n\right) = \left\langle \sum_{k\in I} P_{\mathrm{dir}}(k,n) \right\rangle_M, \tag{10}$$

where $I = \{k \mid \varphi(k,n) = \varphi'\}$, $P_{\mathrm{dir}}(k,n)$ is found with (5), $\langle\cdot\rangle_M$ denotes block averaging over M frames, and φ' is uniformly sampled with L points. The LT-SPD $\Lambda(\varphi', n)$ represents a long-term histogram of all estimated DOAs weighted with the power of the corresponding direct sound. Notice that in (10), only frequency bands k below the spatial aliasing frequency of the array are considered.

Figure 3(a) depicts an exemplary LT-SPD for the case that a sound source (speech source) is active from approximately -80° in a reverberant environment. The higher values in the LT-SPD result from DOA estimates $\widetilde{\varphi}(k,n)$ corresponding to the direct sound (and thus, to the sound source). Due to the temporal averaging in (10), the direct sound forms a larger cluster around the true source position as the sound source possesses a fixed position over time. In contrast, the undesired diffuse sound leads to a specific noise floor in the LT-SPD which is characterized by nearly uniformly distributed random peaks with lower magnitude. It is clear from Figure 3 that this noise floor makes accurate source localisation difficult as the number of sound sources can hardly be estimated. In order to remove this noise floor, we apply at each time instance n of $\Lambda(\varphi', n)$ a dilation filter and erosion filter, both well known from image processing. With these filters, one can remove the noise floor without applying a threshold to the LT-SPD, which usually would be a challenging task. Figure 3(b) depicts the exemplary LT-SPD after applying the dilation (solid line) and erosion (dashed line). The dilation filter, which corresponds to a moving average filter applied along φ', removes smaller gaps in a larger cluster. Subsequently, the erosion filter is applied by setting $\Lambda(\varphi', n)$ at all points φ' to zero if the interval $I = [\varphi' - \Delta\varphi,\ \varphi' + \Delta\varphi]$ contains a point with no power (zero LT-SPD). This removes the thinner clusters (usually corresponding to the diffuse sound power) while maintaining the broader clusters (usually corresponding to the direct sound). Clearly, the erosion filter exploits the fact that diffuse sound leads to a sparse LT-SPD since the DOA estimates $\widetilde{\varphi}(k,n)$ are characterised by a high variance. Therefore, the diffuse sound power appears with narrow peaks at random positions φ'. The required sparsity of the LT-SPD in case of diffuse sound can be assured by choosing a proper angular resolution of $\Lambda(\varphi', n)$, that is, a proper value for L. The optimal L depends on the number of DOA estimates considered for generating $\Lambda(\varphi', n)$ in (10) and on the length M of the temporal block averaging.

(3) *Clustering*. The number $N(n)$ of sound sources and their angular positions $\theta_{1\ldots N}(n)$ are determined by applying a clustering algorithm (similarly to k-means) to the filtered LT-SPD $\Lambda(\varphi', n)$. The used clustering algorithm, in contrast to the traditional k-Means, requires no *a priory* information on the number of sources. It is carried out as follows (cf. Figure 4).

(i) Initial step: generate a vector $\mathbf{v}$ containing Q points from φ' with equal spacing (Q sufficiently large).

(ii) Update step: determine in a limited area around each point in $\mathbf{v}$ the local centre of gravity (COG) in $\Lambda(\varphi', n)$.

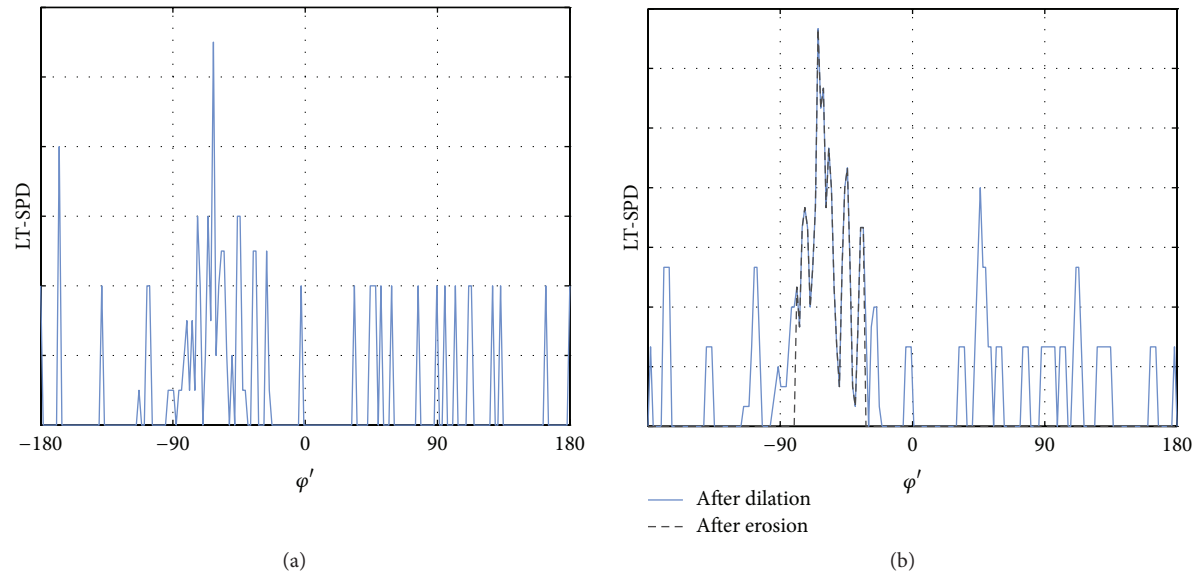


Figure 3: Exemplary LT-SPD when a speaker is active at −80°. (a) Unprocessed LT-SPD. (b) LT-SPD after removing the diffuse sound power.

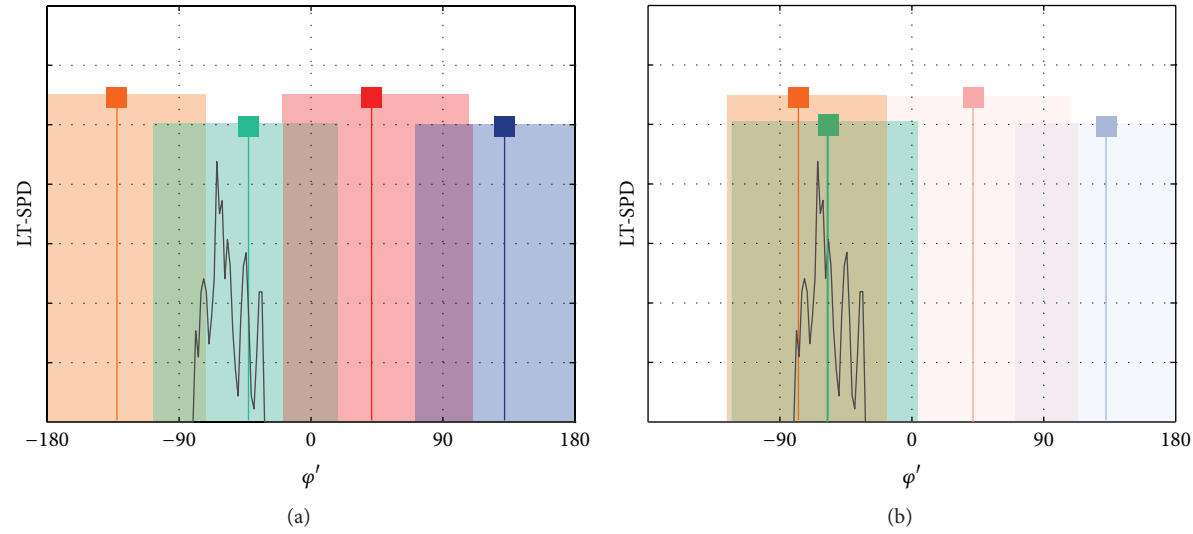


Figure 4: Modified k-means clustering algorithm. (a) Initial step with $Q = 4$. The colored shades represent the areas in which the local COGs are determined. (b) Result after the first assignment step.

(iii) Assignment step: replace the elements in $\mathbf{v}$ by the determined COGs.

(iv) Repeat the update step and assignment step until the stopping criteria (elements in $\mathbf{v}$ remain constant or a specific number of maximum iterations is obtained).

The size of the area around each point in $\mathbf{v}$, for which the COG is computed, is chosen such that the areas of the initial points in $\mathbf{v}$ overlap (see Figure 4(a)). Thus, multiple points in $\mathbf{v}$ might converge to the same position (see Figure 4(b)). In the final step, all points in $\mathbf{v}$ for which $\Lambda(\varphi', n)$ is zero are removed as they likely cover no sound source power. Moreover, identical points or points with close distance are replaced by one average point as they likely cover the same sound source. As result, the remaining points in $\mathbf{v}$ indicate the number $N(n)$ and angular positions $\theta_{1\ldots N}(n)$ of the sound sources.

3.3. Spatial Audio Object Coding. The basic principle behind spatial audio object coding (SAOC) [2] is to represent complex audio scenes by a number of discrete audio object signals. Depending on the application, these audio objects typically comprise single instrumental or vocal tracks (for interactive remixing) or individual speech signals representing the participants in a teleconference. At the receiving side of the SAOC system, the user is allowed to freely mix

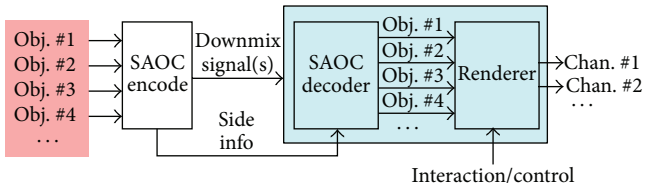


FIGURE 5: Basic structure of SAOC encoding and decoding. The encoder takes separated audio object signals as input; the decoder allows for interactive rendering of the loudspeaker signals.

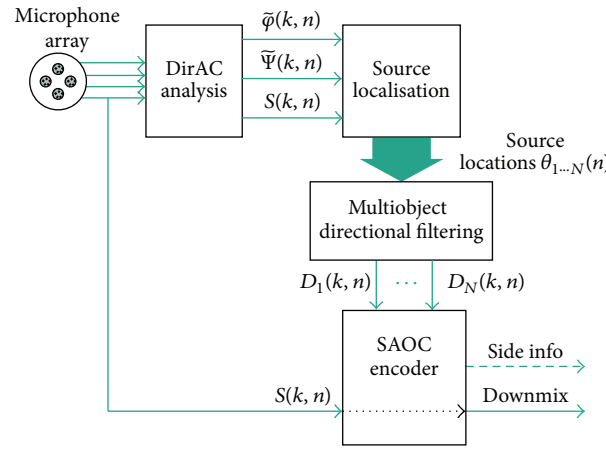


FIGURE 6: Signal processing architecture with DirAC encoding, source localisation, multiobject directional filtering and encoding of the directional filtering, gain functions into SAOC objects. One of the omnidirectional microphone signals is assigned as the downmix signal of SAOC.

the objects according to his/her liking in an interactive way; that is, the level and the position of each audio object may be controlled by the user. Supported playback formats include mono-, stereo, and multi-channel (e.g., ITU 5.1) configurations. In order to save bandwidth, the audio objects are transmitted by means of only one or two downmix audio signals accompanied by parametric side information.

Figure 5 shows the basic structure of the SAOC encoder, the decoder, and the interactive rendering unit. The encoder accepts the individual object signals as input, produces a backward compatible downmix signal, and is responsible for extracting perceptually motivated signal parameters such as object level difference (OLD) and interobject cross Coherence (IOC) in a time/frequency representation [2]. The audio object signals are combined into a mono- or stereo- downmix signal. The parameters describing the downmix process are denoted as downmix gains and transmitted as part of the SAOC side information along with other information such as OLDs and IOCs. This processing results in a compact description of a complex audio scene consisting of a multitude of audio objects, whereas the data rate needed for representing several individual audio objects is significantly reduced down to that required for only one or two downmix channels.

If the objects consist of multiple talkers in the same room, a monodownmix signal $S(k,n)$ can simply be recorded by an omnidirectional microphone. However, each talker's signal has to be separated from the acoustic mixture in order to assign it to an object. This task of acoustic source separation can be efficiently performed in the parameter domain of DirAC, for example, by assigning an instance for directional filtering [19] to each of the N localised acoustic sources.

Directional filtering is based on a short-time spectral attenuation technique and is performed in the spectral domain by a zero-phase gain function, which depends on the estimated instantaneous DOA $\tilde{\varphi}(k,n)$. A so-called directional pattern describes the conversion of the time- and frequency-dependent DOA into a transfer function for each individual time and frequency tile. The directional pattern can be chosen according to the desired application. Directional transfer values close to or equal to one are set for the desired, that is, a source's direction, whereas low transfer values are used for any other direction. In order to separate several talkers from a mixture of sources, several directional filters can be run in parallel. If a given sound scene has to be divided into N objects, N directional filters need to be implemented. Therefore, N gain functions $D_i(k,n)$ are applied to the DirAC omnidirectional signal $S(k,n)$ in parallel, resulting in the separated signal spectra $Y_i(k,n)$ for object i as follows:

$$Y_i(k,n) = S(k,n)\, D_i(k,n). \tag{11}$$

We assume that the original source signals are extracted without loss of energy; that is, we assume that all of the aforementioned downmix gains are one. If there is a diffuse sound, which is not assigned to a localised source and, therefore, not to an audio object, then these sources are represented by a so-called *residual object*, which is represented by individual OLDs and IOCs.

The separated signals $Y_i(k,n)$ may now be processed by an SAOC encoder. As an alternative, it was shown in [20] that the directional filtering gain functions $D_i(k,n)$ can also be transformed into SAOC parameters directly. Some multiplications can be avoided without affecting the separation procedure. Figure 6 shows the efficient structure. The localised sources' angular positions $\theta_{1\ldots N}(n)$ determine the steering of each directional filtering instance. Finally, it should be noted that one of the microphone signals $S_{1\ldots 4}(k,n)$ can directly be assigned to the SAOC downmix signal $S(k,n)$.

4. Semantic Information Extraction

The semantic information is necessary for higher-level stream manipulation and automatic editing, for example, to cut a close-up shot of the person who is currently speaking or to focus on a group of two persons having a dialogue. The corresponding semantic information extraction is performed by several components.

The aim of the face tracking component is to determine at each point in time how many persons are present in the visual scene and where they are in the image. In regard to this higher-level task, the given type of environment, and the required robustness and efficiency of the algorithm, we

propose here to use a method to detect and track the *faces* of persons rather than their full bodies.

The scenario of interest raises a number of challenges for online multiple face tracking:

(1) faces may not be detected for longer periods of time when persons focus on the table or touch screen in front of them (e.g., when playing a distributed game);

(2) when more than two persons are present, they tend to occlude each other more often, leading thus to more frequent track interruptions;

(3) the lighting conditions and scene dynamics are less controlled in a living room environment (than, e.g., in a meeting room);

(4) the assignment of consistent Ids to persons is important for further reasoning and automatic stream editing;

(5) the processing has to be in real time and with a low delay.

The proposed algorithm is an extension of [21] and copes with the previously mentioned challenges in various ways, which will be demonstrated experimentally. Our contributions in this regard are the following:

(1) a state-of-the-art online multiple face tracker in terms of precision and recall over time,

(2) a probabilistic framework for track creation and removal that takes into account long-term observations to cope with false positive and false negative detections [21],

(3) a robust and efficient person reidentification method.

In the following, we will briefly describe the main components of the face tracking system.

4.1. Long-Term Multiple Face Tracking and Person Identification. The proposed tracking algorithm relies on a face detector [22] with models for frontal and profile views. For efficiency reasons, the detector is applied only every 10 frames (i.e., around once per second at a processing speed of around 10 fps). Also, to improve execution speed and reduce false detections, the detector is only scanning image regions with skin-like colours using the discrete model from [23] as a prior and adapting it over time by using the face bounding boxes from the tracker output.

As face detections are intermittent and sometimes rather rare, a tracking algorithm is required. Its goal is to associate detections with tracked objects, to associate tracked objects with persons (person IDs), and to estimate the number and position of visible faces at each point in time. We tackle the tracking problem using a recursive Bayesian framework, where, at each time t, the state X_t is estimated given the observations $Y_{1:t}$ from time 1 to t:

$$p(X_t \mid Y_{1:t}) = \frac{1}{C} p(Y_t \mid X_t) \times \int_{X_{t-1}} p(X_t \mid X_{t-1})\, p(X_{t-1} \mid Y_{1:t-1})\, dX_{t-1}, \tag{12}$$

where C is a normalisation constant. This estimation is implemented using a particle filter with a Markov chain Monte Carlo (MCMC) sampling scheme [4]. The essential components of the particle filter are described in the following (for more details about the MCMC implementation refer to [21]).

(1) *State Space*. We use a multiobject state space formulation, with the global state defined as $X_t = \{X_{i,t}\}_{i=1\cdots M_t}$, where M_t is the number of visible faces at time t. The variable $X_{i,t}$ denotes the state of face i, which comprises the position, scale, and eccentricity (i.e., the ratio between height and width) of the face bounding box.

(2) *State Dynamics*. The overall state dynamics are defined as

$$p(X_t \mid X_{t-1}) \propto p_0(X_t) \prod_{i=1}^{M_t} p(X_{i,t} \mid X_{i,t-1}), \tag{13}$$

that is, the product of an interaction prior p_0 and of the dynamics of each individual visible face. Note that both the creation and deletion of targets are defined outside the filtering step (see next section). The dynamics $p(X_{i,t} \mid X_{i,t-1})$ of visible faces are described by a first-order autoregressive model for the translation components and a first-order model with steady-state for the scale and eccentricity parameters.

The interaction prior p_0 prevents targets to become too close to each other. It is defined between pairs P of visible faces:

$$p(X_t k_t) = \prod_{\{i,j\} \in P} \phi\left(X_{i,t}, X_{j,t}\right) \propto \exp\left\{-\lambda_g \sum_{\{i,j\} \in P} g\left(X_{i,t}, X_{j,t}\right)\right\}, \tag{14}$$

where $g(\cdot)$ is a function penalising overlapping face bounding boxes and λ_g controls the strength of the interaction prior.

(3) *Observation Likelihood*. As a tradeoff between robustness and computational complexity, we employ relatively simple but effective observation likelihood for tracking based on colour distributions. The observation likelihood Y_t is defined as the product of likelihoods of each individual visible face:

$$p(Y_t \mid X_t) = \prod_{i \mid k_{i,t}=1} p(Y_{i,t} \mid X_{i,t}), \tag{15}$$

and the individual observation likelihoods are defined as

$$p\left(Y_{i,t} \mid X_{i,t}\right) \propto \exp\left(-\lambda_D \sum_{r=1}^{6}\left(D^2\left[h^*_{i,t}(r), h\left(r, X_{i,t}\right)\right]\right) - D_0\right), \quad (16)$$

where λ_D and D_0 are constants and $Y_{i,t} = [h(r, X_{i,t})](r = 1 \cdots R)$ are HSV colour histograms computed on different face regions (derived from $X_{i,t}$), at two different quantisation levels, and with decoupled colour and grey-scale bins. $D[\cdot]$ denotes the Bhattacharyya distance between the current observation and the reference histograms $h^*_{i,t}(r)$. The latter are initialised when a new target i is added and adapted slowly over time.

(4) *Target Creation and Removal*. Target candidates are potentially added and removed at each tracking iteration. Traditionally, face detectors have been used to initialise new, targets and targets are removed when the respective likelihood drops. However, face detectors can produce false detections, and, in our scenario, faces may remain undetected for a longer time due to nonfrontal head poses over extended periods. Therefore, we use long-term observations and a probabilistic framework [21] including two Hidden Markov Models (HMM), one helping to decide about track creation and one to decide about removal.

Target Creation. The first HMM estimates the probability of a hidden, binary variable $c_t(i, j)$ indicating at each image position (i, j) if there is a face or not at this position. The posterior probability of c_t can be recursively estimated as

$$p\left(c_t = s \mid O^c_{1:t}\right) = \frac{p\left(O^c_t \mid c_t = s\right) p\left(c_t \mid c_{t-1}\right) p\left(c_{t-1} = s \mid O^c_{1:t-1}\right)}{\sum_{s'} p\left(O^c_t \mid c_t = s'\right) p\left(c_{t-1} = s' \mid O^c_{1:t-1}\right)}, \quad (17)$$

where the transition matrix is defined as $p(c_t \mid c_{t-1}) = 1$ if $c_t = c_{t-1}$, and 0 otherwise. Further, $p(O^c_t \mid c_t) = \prod_{i=1}^{N_c} p(o^c_{t,i} \mid c_t)$, and $o^c_{t,i}$ are the observations. Here, we used two types of observations: the output of a face detector with models for frontal and profile views and a history of previous face positions. The likelihood of the first observation, $p(o^c_{t,1} \mid c_t)$, is defined by the false positive rate and missed detection rate of the face detector; $p(o^c_{t,2} \mid c_t)$ is defined by a parametric model (similar to the one illustrated in Figure 8), that is, a symmetric pair of sigmoid functions (for $c = \{0, 1\}$), the parameters of which are learned beforehand from separate training data (see [21] for more details). Finally, for each detected face that is not associated with any current face target, we compute the following ratio:

$$r^c_t(i, j) = \frac{p\left(c_t(i, j) = 1 \mid O^c_{1:t}(i, j)\right)}{p\left(c_t(i, j) = 0 \mid O^c_{1:t}(i, j)\right)}, \quad (18)$$

at the detection's position (i, j). If $r^c_t > 1$, then a new track is initialised at that position. Otherwise, no track is created.

Target Removal. Decisions on track removal are performed in a similar way, using a second type of HMM. Here, instead of a pixelwise estimation as for creation, the probability of a hidden binary variable $k_{i,t}$ is computed for each tracked target, where $k_{i,t} = 1$ signifies that tracking for target i at time t is correct, and $k_{i,t} = 0$ means that a tracking failure occurred. The decision about removing a target is based on the ratio of posterior probabilities $p(k_{i,t} = K \mid O^k_{1:t})$, where $K = \{0, 1\}$, in analogy to (18), and these posterior probabilities are estimated recursively as in (17). Here, the transition matrix is defined as $p(k_t \mid k_{t-1}) = 0.999$ if $k_t = k_{t-1}$, and 0.001 otherwise. Equally, the observations $O^k_t = [o^k_{t,1}, \ldots, o^k_{t,7}]$ are collected at each time step t and for each target; these observations are the face detections associated with the target, the history of previous face positions, the likelihood of the mean target state, the variance of the target state's position, measures that indicate jumps and drops of the state distribution variance, and a measure that indicates abrupt likelihood drops. The likelihood functions $p(o^k_{t,i} \mid k_t)$ are defined and trained in the same way as for the observations $o^c_{t,i}$ for target creation.

(5) *Person Reidentification*. Whenever the track of a person is lost and reinitialised later or when a person leaves the scene and then comes back, we would like to assign the same identifier (ID) to that person. This is not done inside the tracking algorithm but on a higher level, taking into account longer-term visual appearance observations. More specifically, the person model $P_{j,t}$ of a person j at time t is composed of two colour histograms: a face colour histogram $h^f_{j,t}$ and a shirt colour histogram $h^s_{j,t}$, as well as a long-term history of previous face positions in the image. The structure of the histogram models is the same as the one used for the observation likelihood in the tracking algorithm as described in Section 4.1, that is, two different HSV quantisation levels and decoupled colour and grey-scale bins.

If a target is added to the tracker and there is no existing person model that is unassociated, then a new person model is initialised immediately and associated to the target. Otherwise, the face and shirt colour histograms $h^f_{i,t}$ and $h^s_{i,t}$ of the new target i are computed recursively over r successive frames and stored in $P^*_{i,t}$. After this period, we calculate the likelihood of each stored person model $P_{j,t}$ given an unidentified candidate $P^*_{i,t}$:

$$p\left(P_{j,t} \mid P^*_{i,t}\right) = \exp\left(-\lambda\left(w_f D^2\left[h^f_{j,t}, h^f_{i,t}\right] + w_s D^2\left[h^s_{j,t}, h^s_{i,t}\right]\right)\right) \times p\left(P_{j,t} \mid X_{i,t}\right), \quad (19)$$

where D is the Euclidean distance and the weights are $w_f = 0.25$ and $w_s = 0.75$. The probability $p(P_{j,t} \mid X_{i,t})$ is a distribution over possible identities at the candidate position $X_{i,t}$. This distribution is updated linearly (and normalised) at each time step and for each image position according to the history of tracked target positions. It also contains a small uniform part to allow for reidentification or lost faces that changed their position.

The given person i is then identified by simply determining the person model $P_{m,t}$ with the maximum likelihood:

$$m = \arg\max_{j} p\left(P_{j,t} \mid P^{*}_{i,t}\right), \tag{20}$$

provided that $p(P_{m,t} \mid P^{*}_{i,t})$ is above a threshold. If not, a new person model is created and added to the stored list. All associated person models are updated at each iteration with a small factor $\alpha^{p} = 0.01$. The candidate models are updated with factor $\alpha^{*} = 0.1$.

4.2. Head Pose Estimation and Visual Focus of Attention. Based on the output of the face tracker, the head pose (i.e., rotation in 3 dimensions) of an individual is estimated. The purpose of computing head pose is the estimation of a person's visual focus of attention, which within the context of this work is constrained to being one of the videoconferencing screen, the touch sensitive table, any other person in the room, or "unknown."

Head pose is computed using visual features derived from the 2-dimensional image of a tracked person's head. The features used here are gradient histograms and colour segmentation histograms. The colour segmentation features are estimated from an adaptive Gaussian skin colour model which is used to classify each pixel around the head region as either skin or background, as in [24].

To compensate for the variability in the output of the face tracker, the 2-dimensional face location is reestimated by the head pose tracker. This serves to normalise the bounding box around the face as well as possible while simultaneously using the visual features mentioned previously to estimate pose. This joint estimation of head location and pose improves the overall pose accuracy.

Given the estimated belief (probability distribution) over head pose, the visual focus of attention target is estimated. The range of angles that correspond to each target is modelled using a Gaussian likelihood. The parameters of this Gaussian function (especially the means) are derived from the known spatial locations of the targets within the room. The posterior belief over each target is computed with Bayes' rule using the method of [25].

4.3. Visual Speech and Speaker Detection from Head Motion. Another informative cue is head motion, which will be used in this work to improve the performance of voice activity (i.e., speech) detection. Many existing works proposed to use visual features for speaker detection in videos or other audio-related tasks (e.g., [26–28]). Most of these works attempt to detect people's lip motion. Naturally, this is indeed likely to be an informative visual cue for determining if a person is speaking or not. However, there are several drawbacks with this approach.

(i) Lip motion estimation requires a relatively precise localisation of the mouth region. This is a challenging task when lighting conditions are not controlled, when head pose varies largely, when the (face) image resolution is low, and under motion blur. In some scenarios, the mouth region might not even be visible because of an occlusion (e.g., by the hands) or extreme head pose (e.g., looking down).

(ii) The robust and precise detection of lips in an image is computationally complex in a multiperson, real-time scenario.

To overcome these drawbacks, we make use of the fact that when people speak, they move or behave in a different way. Generally speaking, people who speak move more. Therefore, a relatively simple and efficient visual cue based on the amount of head motion can be used. Here, we leverage the fact that face tracking (described in Section 4.1) provides face regions of the visible persons. From these regions, it is straightforward to efficiently and reliably extract the overall head motion. A more complex model based on full body movements or hand gestures could be considered in the future. However, this could possibly increase the delay for voice activity detection and induce further challenges; for example, in the given scenario, people also move their hands while manipulating the touch screen.

In order to incorporate visual observations over a more extended period of time, that is, not frame-by-frame, we propose a simple Hidden Markov Model (HMM) that estimates the probability of a hidden, binary variable v_t at time t. The value v_t is supposed to be 1 if a person speaks and 0 otherwise. At each time step t and for each person, we estimate the following probability:

$$p\left(v_t \mid o_{1:t}\right) = Z^{-1} \sum_{v'_{t-1}} p\left(o_t \mid v_t\right) p\left(v_t \mid v'_{t-1}\right) p\left(v'_{t-1} \mid o_{1:t}\right), \tag{21}$$

where $o_{1:t}$ are the observations from time 1 to t and Z is a normalisation factor.

Figure 7 illustrates this model. We deliberately modelled v_t for each person independently because we do not want to impose any constraints regarding the interaction of persons at this stage but rather at the audio-visual processing level. The observation o_t is the estimated head motion amount for a given person, that is the mean motion magnitude M inside the face region Ω:

$$o_t = \frac{1}{|\Omega|} \sum_{i \in \Omega} M_t\left(i\right), \tag{22}$$

where at each pixel j of an image, we compute

$$M_t\left(j\right) = \left(1 - \gamma\right) \mathrm{DFD}\left(j\right) + \gamma M_{t-1}\left(j\right), \tag{23}$$

with $\gamma = 0.99$. DFD is the displaced frame difference between the pixel intensities in two successive frames.

The observation likelihood $p(o_t \mid v_t)$ is defined by two symmetric sigmoid functions:

$$\begin{aligned} p\left(o_t \mid v_t = 1, \Theta\right) &= \frac{1}{\pi} \arctan\left(\delta_l\left(o_t - \mu_l\right)\right) + \frac{1}{2} \\ p\left(o_t \mid v_t = 0, \Theta\right) &= 1 - p\left(o_t \mid v_t = 1\right), \end{aligned} \tag{24}$$

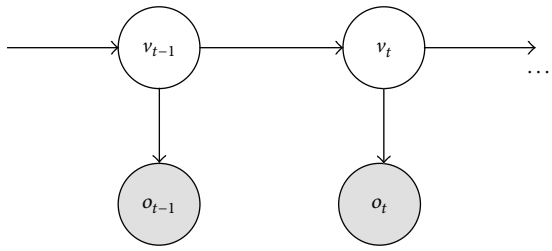


FIGURE 7: The HMM used for each person to estimate voice activity from visual cues. The hidden, binary variable v_t indicates if the person is speaking or not. The probability of v_t is estimated recursively using the previous estimate v_{t-1} and the current observation o_t.

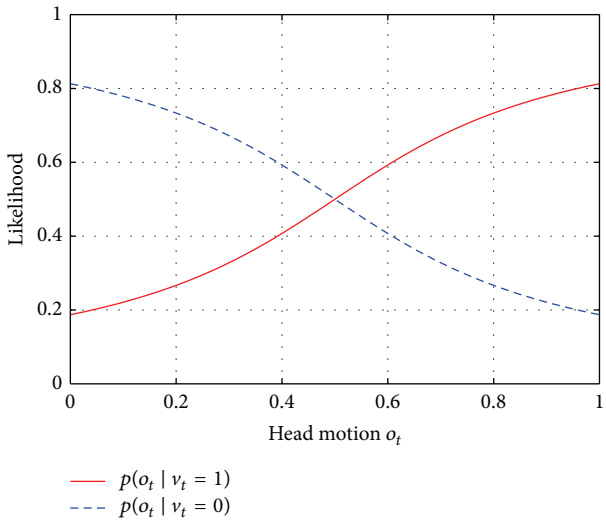


FIGURE 8: Sigmoid functions defining the observation likelihood of head motion for $v_t = 0$ and $v_t = 1$.

where the parameters $\Theta = (\delta_l, \mu_l)$ are determined from separate training data (illustrated in Figure 8). Finally, the posterior probabilities $p(v_t \mid o_t)$ of each person and at each time step t constitute the visual part of the features that is used in multimodal classification experiments. Note that, for simplicity and general applicability, we currently do not train this model for specific persons, and we do not adapt it over time. This could improve the overall results but might also lead to overfitting and drift.

In addition to the speech detection from head motion, the visual-based speaker detection is obtained from the detected speech segments by assigning the relevant person IDs to them.

4.4. Discrete Direction of Arrival Estimation. The instantaneous spatial fingerprints are defined as bit patterns [7] of overlapping sector-based acoustic activity measures, where each sector is represented by one bit of information. The corresponding instances in time refer to processing frames of 32–128 ms length.

Each sector is defined as a 36° wide and 60° high (from the horizontal plane) connected volume of physical space around the microphone array. The sectors are taken in the horizontal plane in steps of 6°. This results in a total of 60 sectors. Wider sectors in smaller steps allow avoiding jittering of acoustic directions and smooth acoustic tracking of dynamic sources.

The sector activity measure is defined as integrated within the sector point-based steered response power with phAse transform weighting (SRP-PHAT). SRP-PHAT [29] in turn can be seen as the sum of generalized cross correlations with phase transform weighting (GCC-PHAT [30]) over all microphone pairs. Further, a sparsity assumption is applied for each frequency bin via minimisation of phase error and the sector activity measures are normalised by the volume of the sector.

Each sector activity measure is thresholded to keep a binary decision, which gives 60 bits of data per each instance in time for a 360° spatial representation. This information is stored as one 64 bit integer value, called the spatial fingerprint.

Finally, this spatial fingerprint is multiplied by the predefined "zone of interest" mask. This multiplication results in directional filtering of the predefined areas of interest, elimination of unnecessary postcalculations, and outlier removal. It can be very helpful in the case of interconnected environments, where audio-visual channels are without an echo suppression mechanism.

The spatiotemporal fingerprint representation is defined as an array of temporally connected spatial fingerprints taken in steps of 16–64 ms. This results in a 2D bit pattern (Figure 9) with a total of 62.5 columns per second and the low bit rate of 500 bytes/second (62.5 long integer values of 64 bits each). The spatiotemporal fingerprints are defined as subsets of the spatiotemporal fingerprint representation (the length depends on the application and can vary from 32 ms to several seconds).

The intersection fingerprint is defined as an intersection in the time domain of all elements within a spatiotemporal fingerprint. Similarly, the union fingerprint is defined as a union in the time domain of all elements within a spatiotemporal fingerprint. The resulting intersection and union fingerprints are normalised at each time instance by keeping single middle "one" out of a group of "ones" per active source.

The intersection fingerprints are used for continuous tracking of acoustic sources by prolonging acoustic trajectories within voice activity segments. The corresponding spatial locations of the active sources are taken from bit positions inside the confirmed intersection fingerprint.

4.5. Voice Activity Detection and Keyword Spotting. Voice activity detection (VAD) covers both verbal and paralinguistic activities and is implemented as a gate. Downstream from the gate, the ASR is unaware that VAD is happening. It just receives segmented data in the same manner as if it was read from a sequence of presegmented utterances. Upstream from the gate, however, the data is actually one continuous stream. The gate segments the input stream in accordance to directional and voice activity/silence information. This can be achieved with an algorithm based on silence models [31] or trained multilayer perceptrons (MLP) using traditional

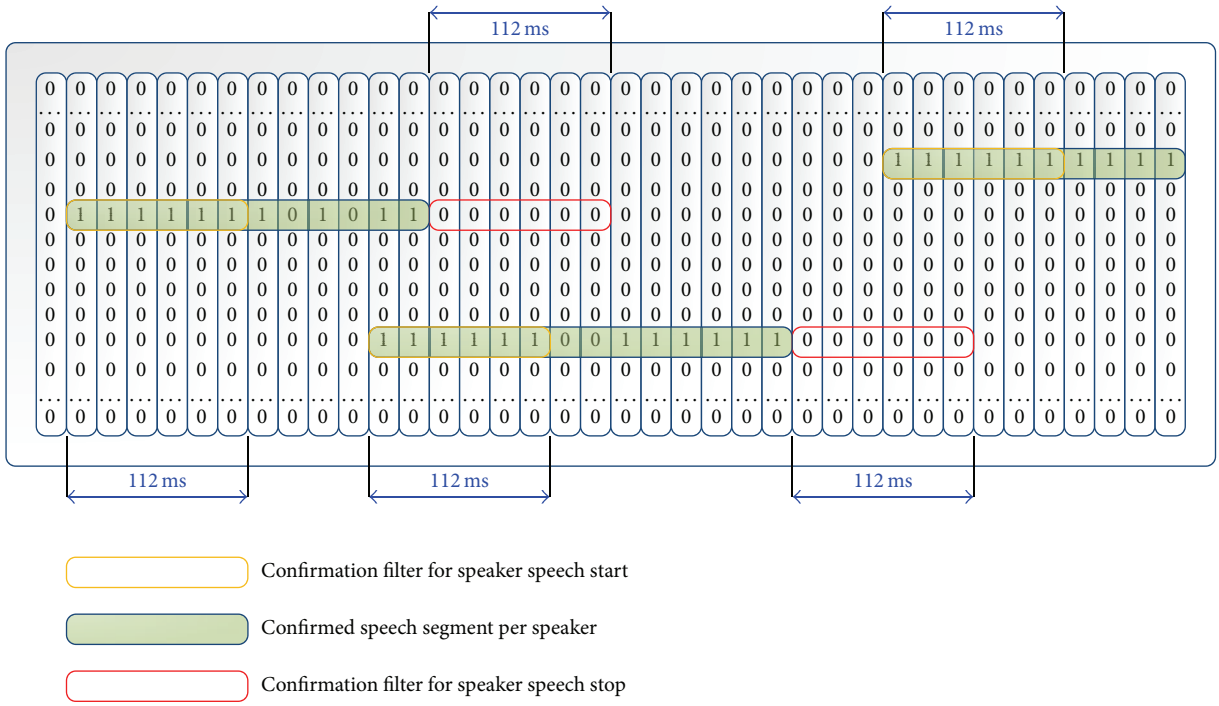


FIGURE 9: Spatiotemporal fingerprint processing. Each column of bits (zeros and ones) represents a spatial fingerprint, a union of several consequent columns represents a spatiotemporal fingerprint. Ones correspond to voice activity; zeros correspond to silence. Horizontal bit position defines instant in time. Vertical bit position defines azimuth with respect to microphone array.

ASR features. However, the current implementation uses adaptively thresholded energy coefficients and directions of arrival to perform localised voice detection. This algorithm works similarly to the traditional VAD module standardized by ETSI for speech coding (AMR1 and AMR2 techniques [32]) and benefits from low complexity and relatively small delay in comparison to more complex VAD techniques, for example, a MLP-based VAD system [33]. The directional information is used to additionally segment voice activity based on a spatial change of the active source position and to filter out the acoustic events, coming from out-of-interest zones.

The ASR component enables speaker-independent large vocabulary-based voice commands and keywords spotting. The spotting is performed based on the predefined list of participants' names and keywords relevant to the given scenario (e.g., orchestrated video chat). In a strict sense, ASR performs the conversion of a speech waveform (as the acoustic realisation of a linguistic expression) into words (as a best decoded sequence of linguistic units). More specifically, the core of the TA2 ASR system is represented by the weighted finite state transducer-(WFST) based token passing decoder known as Juicer [34]. Whilst the decoder is based on a request-driven architecture, the analogues to digital converters (ADCs) are generally interrupt driven. Analysis data flow framework is, in its simplest form, an interface between the decoder's pull architecture and the ADC's push architecture. This framework allows for any directed graph for feature acquisition and is also capable of continuous decoding. Due to the real-time constraints required by the TA2 system, the spotting of keywords is currently performed on 1-best output obtained from the ASR decoder.

4.6. Multimodal Calibration, Association, and Fusion. In our work, we concentrate mainly on score-level fusion and develop a technique [7] which relies on information derived from spatially separated sensors located within a room. Due to the real-time requirements, the association and fusion of person IDs from the video identification with voice activity from the audio channel cannot be delayed until the voice activity is over. The fused events have to be available within a timeframe of two hundred milliseconds to preserve the feeling of instantaneous processing. The low delay temporal association and fusion scheme is depicted in Figure 10.

Audiovisual association is performed between acoustic short-term directional clusters and the positions of tracked faces from the video modality. This involves a mapping estimation between microphone array coordinates (acoustic directional clusters with respect to the microphone array centre) and the coordinates of the image plane, which are defined by the field of view of the camera (Figure 1).

Since the participants do not sit at predefined positions in the room, it can cause ambiguities in the association and fusion. Clearly, the same acoustic short-term directional cluster can correspond to different positions in the image and vice versa. Therefore, the location of a detected face within the image can be mapped to many different sound directions. However, since the participants are mainly located around a table, such ambiguities occur rarely. Therefore, given the mean angle α of the directional cluster from the

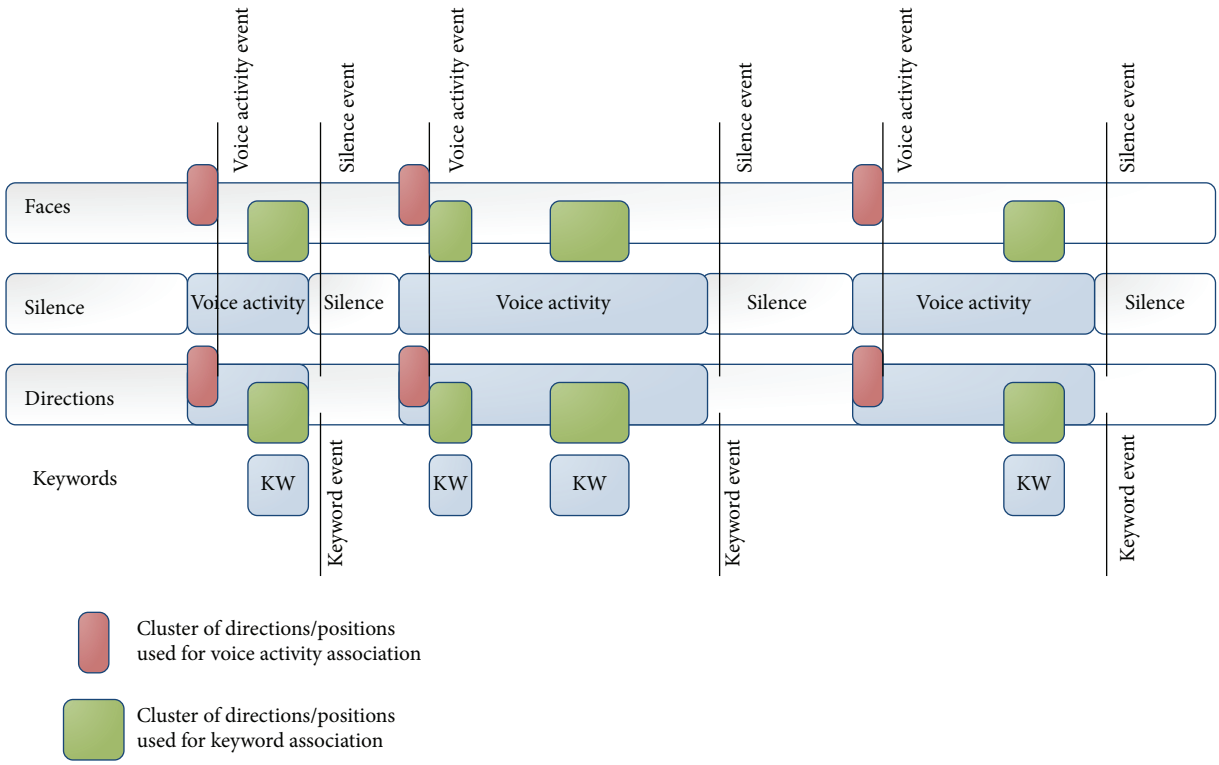


FIGURE 10: Low delay association and fusion. The voice activity is associated with direction of arrival and detected face at the moment of voice activity confirmation and not at the moment when the voice activity is over.

audio modality, a simplified association between a video modality Cartesian coordinate system and audio modality polar coordinate system can be computed as

$$\hat{i} = \arg\min_{i \in P} \left| x_i - x_{\text{ma}} - \gamma \sin \alpha \right|, \tag{25}$$

where P is the set of detected participants from the video modality, x_i is the horizontal position of the ith person, α is the direction of arrival from the audio modality, x_{ma} and γ are calibration parameters: x_{ma} is the horizontal position of the microphone array and γ is the projection weight.

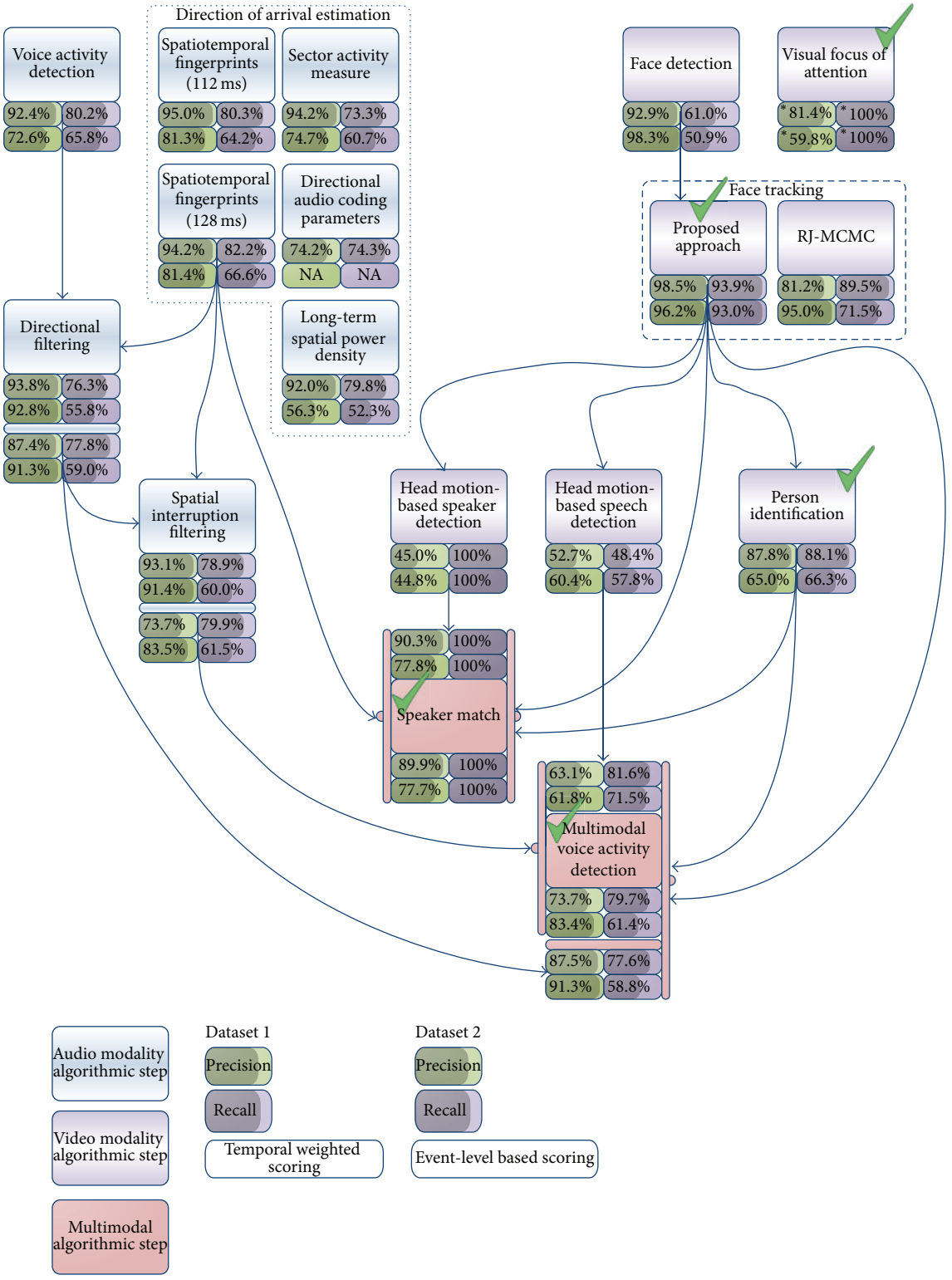
5. Results and Evaluations

5.1. Datasets and Performance Measures. The experiments for objective evaluations were performed on two real life hand-labelled datasets: 3 h 50 min for Dataset 1 with enabled echo suppression [35] (the process of removing echo from a voice communication in order to improve voice quality on a teleconferencing call); 1 h 20 min for Dataset 2 [36] with disabled echo suppression, lower SNR, and fewer frontal face views. Dataset 2 was made publicly available. The datasets follow the systematic description presented in [36] and contain 2 room recorded gaming sessions with enabled video chat of socially connected but spatially separated people. Each room was recorded and analysed separately and contained up to 4 people.

The achieved results at different steps of processing are summarised in Figure 11. Precision is defined as the number of true positive test events (test events correctly detected as belonging to the positive class) divided by the total number of test events detected as belonging to the positive class (the sum of true positive and false positive test events). Recall is defined as the number of true positive test events divided by the total number of test events that actually belong to the positive class (the sum of true positive and false negative test events). In addition to event-level based scoring, we consider temporal weighted scoring to better evaluate algorithms from the perspective of amount of time. In case of temporal weighted scoring, precision is defined as the total time of true positive test events (test events correctly detected as belonging to the positive class) divided by the total time of test events detected as belonging to the positive class (the sum of true positive and false positive test segments). Recall is defined as the total time of true positive test events divided by the total time of test events that actually belongs to the positive class (the sum of true positive and false negative test events).

Achieved results presented in Figure 11, mostly given in terms of precision and recall, should rather be seen as complementary (more rigorous results are presented in the other figures). Since the individual processing blocks were evaluated with locally selected operating points, both precision and recall, were varying in the different steps of the evaluations.

5.2. Face Tracking and VFOA Results. The block "face detection" (see Figure 11) shows the precision and recall of a standard face detector, described in [22], computed



*Evaluations for visual focus of attention are based on subset of datasets (30 min for Dataset 1, 5 min for Dataset 2)

Figure 11: Evaluations at different steps of processing. Two different scorings of results are used: (a) temporal weighted scoring is used to better evaluate algorithms from the perspective of amount of time. (b) Event-level based scoring is used to better evaluate algorithms from the perspective of amount of discrete events. Upper blocks show basic precision/recall values; further blocks show achieved precision/recall values after each step of processing for the operating system's point. The two lowest blocks (speaker match and multimodal voice activity detection) show the final achieved precision/recall values. The results from the blocks, marked with a tick, are propagated further to the TA2 system, while the results from other blocks are intermediate or comparative.

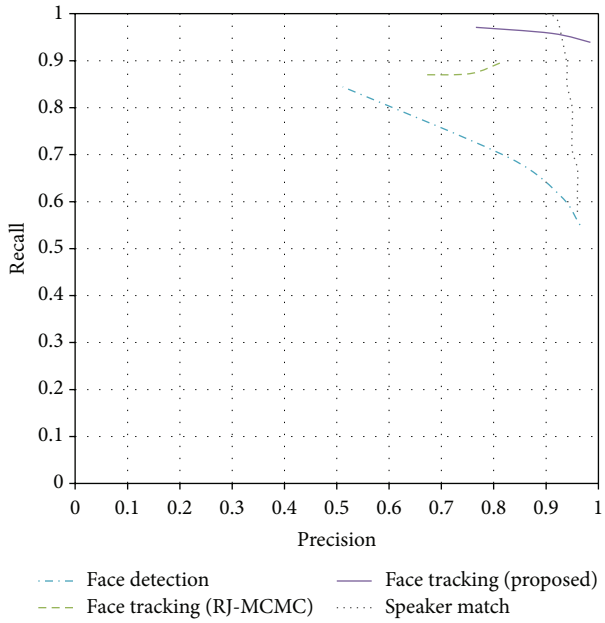


Figure 12: Recall versus precision for face detection, face tracking, and speaker match (Dataset 1). Both face tracking and speaker match show good performance as there are only two participants within a sector of 100°.

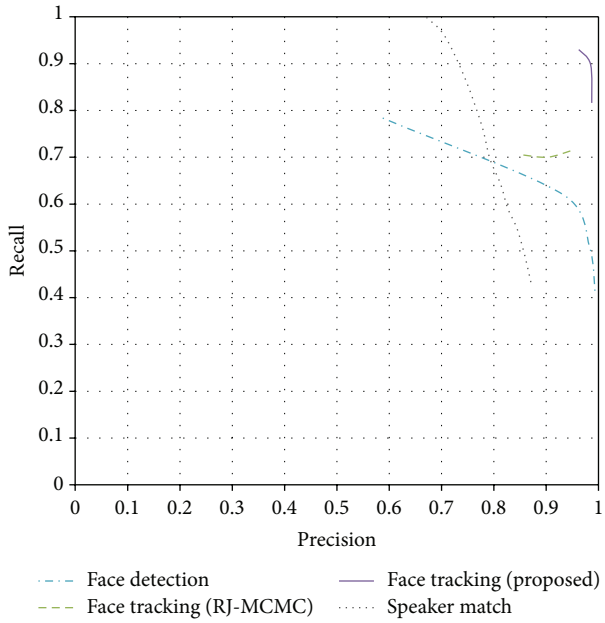


Figure 13: Recall versus precision for face detection, face tracking, and speaker match (Dataset 2). Speaker match shows lower performance in case of Dataset 2 due to the presence of 4 participants within a sector of 100°.

as the average over all people. The block "face tracking", shows the results of the face tracking algorithm, described in Section 4.1, which improves the overall accuracy of the video processing. The corresponding dependencies between recall and precision are shown in Figures 12 and 13. It is clearly visible that the proposed approach for face tracking outperforms both the standard face detector [22] and the RJ-MCMC method [4]. More extensive face tracking evaluations are presented in [21], where we have shown that the recall is increased by relative 7.8% while the false positive rate is decreased by relative 38.3% compared to a state-of-the-art multiple target tracking algorithm [4]. In addition to face tracking, the person identification algorithm (described in detail in Section 4.1) has been evaluated on the given datasets by measuring the amount of time with correctly and incorrectly assigned identifiers, respectively, where, for a given person, the longest continuous track determines the correct identifier. Then, precision and recall, shown in Figure 11, are computed in a standard way. We also performed a visual focus of attention (VFOA) evaluation (see Figure 11) using a representative subset of the data, where we manually annotated for each frame and each person (if not ambiguous) whether the person is looking at table, screen, another person (ID) or none of them (unfocused). Nonannotated, ambiguous frames were not included in the statistics.

5.3. Speaker Match Results. The speaker match is evaluated (i.e., temporal weighted scoring) based on different acoustic localisation approaches (see Figure 11), described in Sections 3.2 and 4.4. In case of the spatiotemporal fingerprints approach for speaker match [7], defined by block "spatiotemporal Fingerprints" in Figure 11, the dependency between recall and precision for varying operating point is shown in Figures 12 and 13. Here, the fingerprint approach with algorithmic delay of about 112 ms is visualised. From these figures, it is clearly visible that for Dataset 1, the speaker match performs significantly better than for Dataset 2 since there are 4 participants in Dataset 2 within a sector of 100°, which is definitely going beyond the spatial resolution of the used microphone array. We also assume that the speaker match approach based on spatiotemporal fingerprints [7] suits better the task of discrete semantic event extraction, while the approach based on long-term spatial power density suits better spatial audio object coding (SAOC) [2] as it allows continuous tracking of the audio object (see Figures 14 and 15). Achieved results of spatiotemporal fingerprints, shown in Figure 11, are also compared to sector activity measure [37] and directional audio coding techniques [15] (evaluated in the same manner).

In addition to the temporal weighted scoring hitherto presented, we also performed an event-level based scoring defined by block "speaker match" in Figure 11. In this case, an event represented by a speech segment needs to be assigned with detected speaker face. Since the task is not detection but rather identification (of a speaker), the performance is measured in terms of accuracy (variable precision with a fixed recall of 100%). In the simplest case, the speaker match is based on mapping of direction of arrival to a corresponding detected face (using (25)). Achieved speaker (localisation) match accuracies are about 89.9% and 77.7% for Dataset 1 and Dataset 2, respectively.

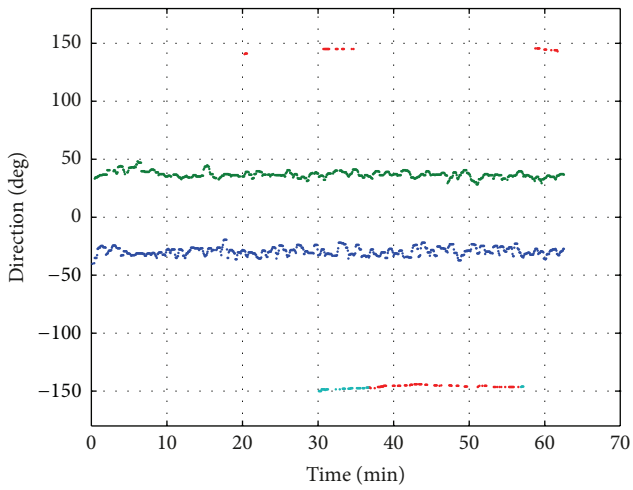


Figure 14: Visualisation of continuous audio speech tracking for a subsession of 1 h 02 min (Dataset 1) performed using long-term spatial power density algorithm. The tracks are assigned to 2 different participants between −50° and 50°. Additional track at around 180° corresponds to the remaining artefacts from remote echo cancellation.

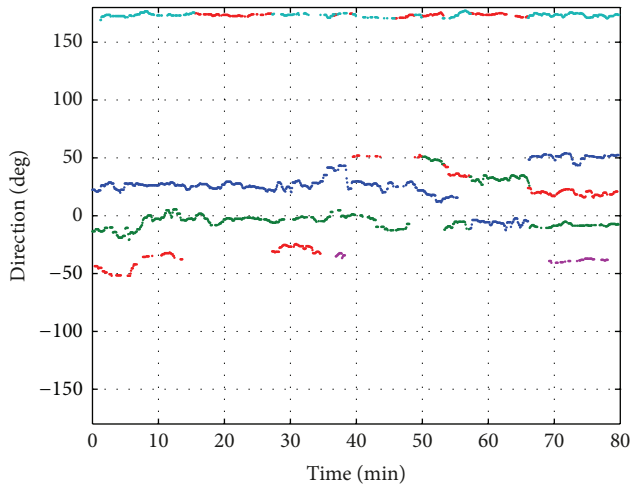


Figure 15: Visualisation of continuous audio speech tracking for a session of 1 h 20 min (Dataset 2) performed using long-term spatial power density algorithm. The tracks are assigned to 4 different participants between −50° and 50°. Additional track at 180° corresponds to remote echo.

We have also carried out event-level based speaker match experiments by exploiting purely information extracted by a visual head motion analysis (see Section 4.3). This is defined by block "head motion based speaker detection" in Figure 11. The mean, estimated over the given speech interval, represents a confidence value of visual head motion for each individual speaker. The maximum over the mean estimates determines the recognised (localised) speaker in the given speech interval. Using this technique, the accuracies of about 45.0% and 44.8% are achieved for Dataset 1 and Dataset 2, respectively.

Eventually, we performed an audio-visual combination of independent streams to possibly improve speaker localisation (defined also in block "Speaker Match" in Figure 11). A relatively simple, scenario-independent and real-time linear combination of audio and visual streams was performed, where the current speaker $\widehat{i}$ is determined by (25).

As weighting factors, a normalised distance was taken for audio stream (estimated by the previous equation, where argmin operation is removed). In the equation, P is the set of detected participants from the video modality, x_i is the horizontal position of ith person, α is the direction of arrival from the audio modality, and x_{ma} and γ are calibration parameters: x_{ma} is the horizontal position of the microphone array γ is the projection weight. In the video modality, the mean confidences of visual head motion were exploited. These weights were furthermore modified by a prior which rather takes into account the audio stream against the video stream.

Results, given in the block "speaker match" in Figure 11, obtained after audio-visual combination show slight additional improvements (absolute accuracies of 90.3% and 77.8% for Dataset 1 and Dataset 2, resp.) over the audio-only system (absolute accuracies of 89.9% and 77.7% for Dataset 1 and Dataset 2, resp.), as shown in Figure 11. According to preliminary experiments on other additional data, we have discovered that the gain achieved by augmenting the visual information (i.e., head motion estimation) is more significant in case of more noisy audio data.

Known meeting-wise speaker error rates for CPU-intensive state-of-the-art speaker diarisation techniques [38] are as low as 7.0% for realigned MFCC+TDOA combination of the HMM/GMM system with optimal weights and for Kullback-Leibler-based realigned MFCC+TDOA combination of the information bottleneck system with optimal weights. In the case of automatic weights, overall speaker error rates are about 13% and 10% correspondingly. These state-of-the-art estimates are given only as an overview and cannot be used for direct comparison as the data, hardware and scenario used in our experiments differ from the data, hardware and scenario used in [38]. In addition, the state-of-the-art systems have a latency of 500 ms and a state of minimum 3 seconds duration, while we were able to achieve reasonably good results with an algorithmic delay and minimum state duration as low as 128 ms, which is more crucial for TA2 scenarios. We should note that the algorithmic delay does not include capturing delay, which in turn can result in additional 10–20 ms. Naturally, there is a tradeoff between lower latency and better accuracies. Systems that are not requiring the lowest possible delay can potentially achieve higher accuracies.

5.4. Voice Activity Results. The block "voice activity detection" and derivative blocks (Figure 11) show precision and recall for the operating system's point performed on the output of the local far-field voice activity detection (more than 6 K manually annotated speech segments used). Although only Dataset 1 is echo cancelled, we were able to achieve reasonably good precision/recall levels for Dataset 2 (see

Figure 11) after application of the "Directional filtering" block on semantic level within voice activity detector (a difference of 20.2% in precision (92.8% instead of 72.6%) can be seen between corresponding blocks). The sector of interest in the final system for directional filtering was defined as [−110°, 110°] with respect to the reference direction of 0°, defined as an imaginary arrow intersecting the camera and the centre of the microphone array, facing the participants. This allows us to eliminate remote parties in case of disabled echo suppression (Dataset 2) and few echo cancelation artefacts in case of enabled echo suppression (Dataset 1).

The block "directional filtering" shows precision and recall values of voice activity detection for the case when barge-in (break into a conversation) events are treated by temporal interruption detector, while the blocks "spatial interruption filtering" show precision and recall values of voice activity detection for the case when barge-in events are treated by a spatial interruption detector (i.e., using azimuth of the stream). While the approach with spatial interruption detection shows slightly better performance using temporal weighted scoring, surprisingly, we have found that in case of event-level based scoring, the spatial approach has a lower performance. We presume that this is due to some false alarms being fragmented into shorter ones.

In addition to the audio modality, "head motion based speech detection" given in Figure 11, exploiting purely information extracted by a visual head motion analysis (see Section 4.3) is evaluated for the operating system's point. The event-level based performance of voice activity detection (VAD) based on fusion of multimodal information is represented by the "multimodal voice activity detection" block in Figure 11. The performance is influenced by Face Detection and Person Identification algorithms due to assigning the generated voice activity segments to a visually tracked person. Besides using the audio modality only to generate the events (i.e., speech segments generated by ACDE), we also perform the subsequent fusion of these audio events with visual events estimated by head motion-based speech detection algorithm (performed in VCDE) to improve the overall VAD performance. More specifically, the "multimodal voice activity detection" block in Figure 11 compares 3 systems evaluated for the operating system's point: (a) complete Multimodal VAD; (b) and (c) VAD relying only on energy-based audio estimates (no head motion employed here) with and without applying the block of spatial interruption filtering, respectively.

We realise that the evaluation using precisions and recalls for an operating selected by the system is not informative enough, since the numbers among different blocks, as presented in Figure 11, cannot be directly compared. Therefore, in addition, the VAD performance is also evaluated by employing detection error tradeoff (DET) curves of miss versus false alarm probabilities evaluating the detection for a large set of operating points [39]. These probabilities are estimated using the absolute number of targets (i.e., the number of speech segments comprised in the transcription) as well as nontargets (i.e., the number of potential speech segments not comprised in the transcription but appearing in the detection output). The resulting DET curves are normalised in such a way that the number of targets and nontargets is set to be equal. For each operating point in DET curve, precision and recall values can be estimated. Thus, depending on a potential application, VAD can easily be tuned by considering different thresholds applied on confidence scores associated with each speech segment.

Figures 16 and 17 show DET characteristics for detection of voice activity on Datasets 1 and 2. More specifically, 5 different audio-visual VAD systems were considered based on the input audio and a visual motion extracted from video stream.

(i) Audio: the events (i.e., speech segments) are purely detected from the audio signal in the block of ACDE, together with confidence scores. This corresponds to system (b) hitherto presented in the block of multimodal voice activity detection.

(ii) Video: the events (i.e., speech segments) are purely detected from the video using head motion-based speech detection algorithm (described in detail in Section 4.3).

(iii) Audio + video no. 1: the events (i.e., speech segments) are detected from both modalities and are merged in case of their overlap; the confidence scores from audio and video are linearly weighted. This corresponds to system (a) hitherto presented in the block of multimodal voice activity detection.

(iv) Audio + Video no. 2: the events (i.e., speech segments) are detected from audio only, however the corresponding confidence scores are estimated using the visual motion algorithm.

(v) Audio + Video no. 3: the events (i.e., speech segments) are detected from audio only; however the assigned confidences are given by the combination of acoustic and visual confidence scores.

Graphical outputs presented by the DET plots in Figures 16 and 17 indicate that the VAD based on both audio and video modalities (audio + video no. 1) outperforms audio-only VAD for most of the potential operating points. In more detailed view, the largest improvement was obtained for audio + video no. 1 VAD system, where the events (i.e., speech segments) are first detected independently from both modalities and then merged into a single output stream of events. In case of the simple scenario provided by Dataset 1, where the audio signals from the remote rooms were well separated using echo cancellation and the audio has relatively high SNR, the audio + video combination did not significantly improve over audio-only VAD. It can be seen that audio-based VAD outperforms video-based VAD. However, the combination of Audio and Video is able to enlarge a potential set of operating points (especially when a low false alarm rate is expected). In Dataset 2 the audio is not echo cancelled, and combined Audio + Video offers better detection results over the whole DET curve (especially for low miss probabilities) compared to uni-modal VAD systems.

5.5. SAOC Results. The transcoding of separated audio objects (using DirAC-based directional filtering [15]) to

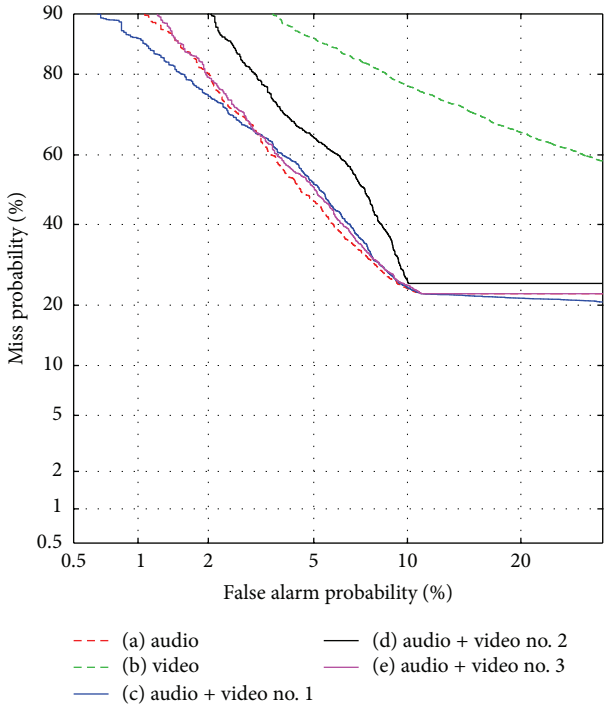


Figure 16: DET plot of voice activity detection performance for Dataset 1: solid lines—audio + video combinations, dashed lines—audio and video systems individually. VAD based on both audio and video modalities (audio + video no. 1) indicates better performance than audio-only VAD for most of the operating points.

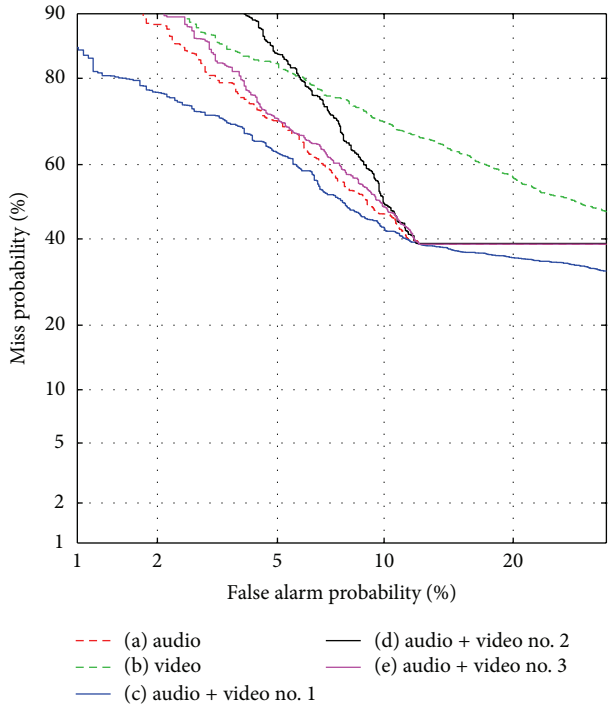


Figure 17: DET plot of voice activity detection performance for Dataset 2: solid lines—audio + video combinations, dashed lines—audio and video systems individually. VAD based on both audio and video modalities (audio + video no. 1) indicates better performance than audio-only VAD.

SAOC objects [2] has to be evaluated with respect to a negligible loss of quality compared to other parametric spatial coding techniques. If we achieve comparable audio quality, SAOC offers the desired advantage of extensive user interaction. In [20], SAOC has been compared against DirAC on the basis of a MUSHRA listening test [40]. Both coding techniques SAOC and DirAC were based on a single-channel downmix signal. An uncoded stereo signal, namely, an M/S-stereo signal served as a reference. A monodownmix of the M/S-stereo signal served as a lower anchor.

The recorded microphone signals were provided as B-format, comprising an omni-directional signal W and dipole signals X and Y. The omni-directional and the dipole signals were used for the M/S-stereo reference signal. Six test items were recorded using a multichannel loudspeaker playback setup in a mildly reverberant room. The sound scenes consisted of either two or three talkers arranged at $+60^{\circ}$, -60° and 0° and incorporated single and double talk situations. For three items, diffuse background noise (recorded at a trade show) was added with an SNR of 9 dB.

In addition to the reference M/S-stereo signal, we encoded the B-format signal into DirAC and directly rendered it to a conventional stereo setup. Test systems StrfFwd (SAOC) and Efficient (efficient DirAC-to-SAOC) included transcoding from DirAC to SAOC. Depending on the number of active talkers, two or three directional filtering instances were steered towards the sources (loudspeakers). For system StrfFwd, we calculated separated source signals prior to SAOC encoding; system Efficient resulted from direct efficient transcoding from directional filtering to SAOC objects [20]. The mono anchor represented system LowAnchor.

Figure 18 shows the results from the MUSHRA listening test (with respect to a negligible loss of quality compared to other parametric spatial coding techniques). The reference system could clearly be distinguished from the coded systems. Evaluation was mainly based in the spatial image, which slightly differed using SAOC. No coding artefacts or timbral colorations have been reported by the eight expert listeners. Therefore, the DirAC-to-SAOC transcoding scheme can be rated as only slightly inferior to the DirAC system. It should be noted that only SAOC offers the advantage of a large degree of user interactivity.

5.6. Computational Cost Analysis. The system architecture is grouped into 4 main parts, as illustrated in Figure 2. The current implementation assumes that each of these 4 parts is running on an individual CPU core of a 64-bit PC to meet the real-time constraints of the whole system. More specifically, we use a TCP socket implementation to detach the ACE block (providing the echo-cancelled audio recordings from the microphone array) from the other blocks. The ACE directly communicates with the ACDE which is installed with the rest of blocks (VCDE and UCDE) on a 4-core CPU (i.e., Intel(R) Core (TM) i7 CPU at 2.8 GHz 12 GB RAM).

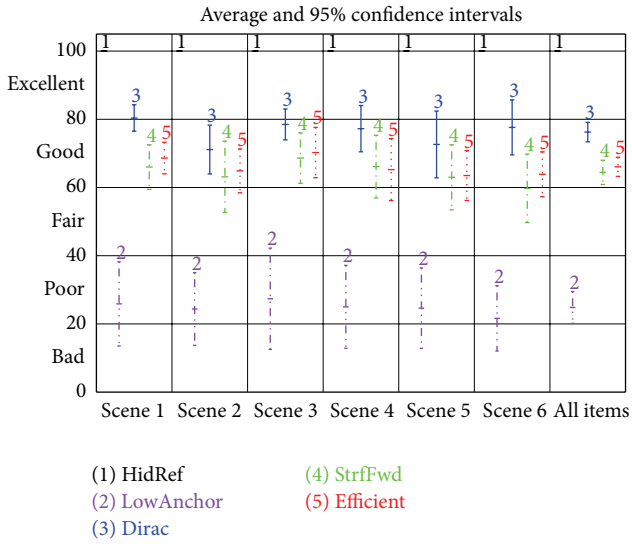


Figure 18: Results of a MUSHRA listening test comparing DirAC (Dirac) against a sequence of directional filtering, SAOC (StrfFwd) and efficient DirAC-to-SAOC transcoding (Efficient). A M/S stereo signal served as a reference (HidRef), while a monodownmix represented the lower anchor (LowAnchor).

ACE, VCDE, and UCDE can operate approximately 10 times in real time. The most complex part is ACDE which contains a large vocabulary continuous speech recogniser. The real-time performance of ACDE is controlled by optimising the decoder parameters (i.e., pruning).

6. Conclusion

In this paper, we presented a system aimed at enabling higher-level multimodal stream manipulation, while fulfilling the specific requirements of the TA2 scenario and addressing the corresponding challenges: streams and semantic information need to be computed in real time with low delay from spatially separated sensors (within a room) in an open, unconstrained environment; the system tracks a potentially varying number of persons who are not constrained to sit at specific places; the detected events need to be reliably and consistently associated to the involved people.

More particularly, an intelligent audio capturing block transforming the input sound into individual acoustic objects was developed to be applied in reverberant environment. Such acoustic objects representing an analysed sound scene can consist of multiple speech sources appearing in the same room recorded by an omnidirectional microphone array. The audio source localisation is then performed using a power-weighted histogram of the DOA estimates corresponding to the directional sound followed by the clustering algorithm providing the final number of sound sources and their positions. Finally, an object-based representation using MPEG-SAOC is used for transmission.

For higher-level stream manipulation, the semantic information extraction is performed using various components from audio-visual input. The visual information is exploited in face tracking, person identification, head pose estimation, visual focus of attention, and visual speech, and speaker detection components. Audio input provided by SAOC is used in direction of arrival, voice activity detection and keyword spotting components. Eventually, audio-visual association and fusion is performed to generate bimodal cue estimates to be exploited in the subsequent higher-level processing.

Overall, our main contributions are the design of an integrated real-time system with latency below 130 ms comprising several state-of-the-art audio-visual processing algorithms and a thorough performance evaluation of the different components of the system on two different challenging datasets. The main evaluated components of the system are face tracking, speaker localisation and match, multimodal voice activity detection, estimation of visual focus of attention, and spatial audio object coding with respect to a negligible loss of quality compared to other parametric spatial coding techniques.

Acknowledgment

The research leading to these results has received funding from the European Community's 7th Framework Programme ICT Integrating Projects TA2 (Grant agreement no. 214793).

References

[1] M. Falelakis, R. Kaiser, W. Weiss, and M. F. Ursu, "Reasoning for video-mediated group communication," in *Proceedings of the 12th IEEE International Conference on Multimedia and Expo (ICME '11)*, Barcelona, Spain, July 2011.

[2] J. Engdegård, B. Resch, C. Falch et al., "Spatial audio object coding (SAOC)—the upcoming MPEG standard on parametric object based audio coding," in *Proceedings of the 124th AES Convention*, Amsterdam, The Netherlands, 2008.

[3] J. Carletta, S. Ashby, S. Bourban et al., "The AMI meeting corpus," in *Proceedings of the Machine Learning for Multimodal Interaction (MLMI '05)*, Edinburgh, UK, 2005.

[4] Z. Khan, T. Balch, and F. Dellaert, "MCMC-based particle filtering for tracking a variable number of interacting targets," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 11, pp. 1805–1819, 2005.

[5] J. Ajmera, *Robust audio segmentation [Ph.D. thesis]*, Ecole Polytechnique Federale de Lausanne (EPFL), 2004.

[6] J. Pardo, X. Anguera, and C. Wooters, "Speaker diarization for multi-microphone meetings using only between-channel differences," in *Proceedings of the Machine Learning for Multimodal Interaction (MLMI '06)*, Bethesda, Md, USA, 2006.

[7] D. Korchagin, "Audio spatio-temporal fingerprints for cloudless real-time hands-free diarization on mobile devices," in *Proceedings of the 3rd Joint Workshop on Hands-free Speech Communication and Microphone Arrays (HSCMA '11)*, pp. 25–30, Edinburgh, UK, June 2011.

[8] C. Sanderson and K. K. Paliwal, "Information fusion and person verification using speech and face information," Idiap Research Report IDIAP-RR 02-33, 2002.

[9] M. Slaney and M. Covell, "Facesync: a linear operator for measuring synchronization of video facial images and audio

tracks," in *Proceedings of the Neural Information Processing Systems*, pp. 814–820, 2000.

[10] D. Korchagin, P. Motlicek, S. Duffner, and H. Bourlard, "Just-in-time multimodal association and fusion from home entertainment," in *Proceedings of the 12th IEEE International Conference on Multimedia and Expo (ICME '11)*, Barcelona, Spain, July 2011.

[11] J. Hershey and J. Movellan, "Audio vision: using audio-visual synchrony to locate sounds," in *Proceedings of the Neural Information Processing Systems*, pp. 813–819, 1999.

[12] H. Nock, G. Iyengar, and C. Neti, "Speaker localisation using audio-visual synchrony: an empirical study," in *Proceedings of the 2nd International Conference on Image and Video Retrieval (CIVR '03)*, Urbana-Champaign, Ill, USA, 2003.

[13] M. Gurban and J. Thiran, "Multimodal speaker localization in a probabilistic framework," in *Proceedings of the European Signal Processing Conference (EUSIPCO '06)*, Florence, Italy, 2006.

[14] K. Otsuka, S. Araki, K. Ishizuka, M. Fujimoto, M. Heinrich, and J. Yamato, "A realtime multimodal system for analyzing group meetings by combining face pose tracking and s peaker diarization," in *Proceedings of the 10th International Conference on Multimodal Interfaces (ICMI '08)*, pp. 257–264, Chania, Greece, October 2008.

[15] V. Pulkki, "Spatial sound reproduction with directional audio coding," *Journal of the Audio Engineering Society*, vol. 55, no. 6, pp. 503–516, 2007.

[16] S. Rickard and Ö. Yilmaz, "On the approximate W-disjoint orthogonality of speech," in *Proceedings of the IEEE International Conference on Acustics, Speech, and Signal Processing (ICASSP '02)*, pp. I/529–I/532, May 2002.

[17] O. Thiergart, G. Del Galdo, M. Prus, and F. Kuech, "Three-dimensional sound field analysis with directional audio coding based on signal adaptive parameter estimators," in *Proceedings of the AES 40th International Conference on Spatial Audio: Sense the Sound of Space*, Tokyo, Japan, October 2010.

[18] O. Thiergart, R. Schultz-Amling, G. Del Galdo, D. Mahne, and F. Kuech, "Localization of sound sources in reverberant environments based on directional audio coding parameters," in *Proceedings of the 127th AES Convention*, New York, NY, USA, 2009.

[19] M. Kallinger, H. Ochsenfeld, G. Del Galdo et al., "A spatial filtering approach for directional audio coding," in *Proceedings of the 126th AES Convention*, 2009.

[20] J. Herre, C. Falch, D. Mahne, G. Del Galdo, M. Kallinger, and O. Thiergart, "Interactive teleconferencing combining spatial audio object coding and DirAC technology," in *Proceedings of the 128th AES Convention*, London, UK, 2010.

[21] S. Duffner and J.-M. Odobez, "A track creation and deletion framework for long-term online multi-face tracking," *IEEE Transactions on Image Processing*, vol. 22, no. 1, pp. 272–285, 2013.

[22] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 1, pp. 511–518, December 2001.

[23] C. Scheffler and J. M. Odobez, "Joint adaptive colour modelling and skin, hair and clothing segmentation using coherent probabilistic index maps," in *Proceedings of the British Machine Vision Conference*, 2011.

[24] E. Ricci and J.-M. Odobez, "Learning large margin likelihoods for realtime head pose tracking," in *Proceedings of the IEEE International Conference on Image Processing (ICIP '09)*, pp. 2593–2596, November 2009.

[25] S. O. Ba and J.-M. Odobez, "Recognizing visual focus of attention from head pose in natural meetings," *IEEE Transactions on Systems, Man, and Cybernetics B*, vol. 39, no. 1, pp. 16–33, 2009.

[26] D. Sodoyer, B. Rivet, L. Girin, J.-L. Schwartz, and C. Jutten, "An analysis of visual speech information applied to voice activity detection," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '06)*, pp. I601–I604, May 2006.

[27] S. Siatras, N. Nikolaidis, M. Krinidis, and I. Pitas, "Visual lip activity detection and speaker detection using mouth region intensities," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 19, no. 1, pp. 133–137, 2009.

[28] H. Hung and S. O. Ba, "Speech/non-speech detection in meetings from automatically extracted low resolution visual features," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '10)*, Dallas, Tex, USA, 2010.

[29] 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, pp. 320–327, 1976.

[30] J. DiBiase, H. Silverman, and M. Brandstein, "Robust localization in reverberant rooms," in *Microphone Arrays*, M. Brandstein and D. Ward, Eds., chapter 8, Springer, 2001.

[31] P. N. Garner, "Silence models in weighted finite-state transducers," in *Proceedings of the 9th Annual Conference of the International Speech Communication Association (INTERSPEECH '08)*, pp. 1817–1820, Brisbane, Australia, September 2008.

[32] GSM 06. 94, Digital cellular telecommunications system (Phase 2+), "Voice activity detector (VAD) for adaptive multi rate (AMR) speech traffic channels," 1999.

[33] J. Dines, J. Vepa, and T. Hain, "The segmentation of multi-channel meeting recordings for automatic speech recognition," in *Proceedings of the INTERSPEECH and 9th International Conference on Spoken Language Processing (INTERSPEECH ICSLP '06)*, pp. 1213–1216, September 2006.

[34] P. N. Garner, J. Dines, T. Hain et al., "Real-time ASR from meetings," in *Proceedings of the 10th Annual Conference of the International Speech Communication Association (INTERSPEECH '09)*, pp. 2119–2122, Brighton, UK, September 2009.

[35] F. Kuech, M. Kallinger, M. Schmidt, C. Faller, and A. Favrot, "Acoustic echo suppression based on separation of stationary and non-stationary echo components," in *Proceedings of the Acoustic Echo and Noise Control*, Seattle, Wash, USA, 2008.

[36] S. Duffner, P. Motlicek, and D. Korchagin, "The TA2 database: a multimodal database from home entertainment," in *Proceedings of the Signal Acquisition and Processing*, Singapore, 2011.

[37] G. Lathoud and I. A. McCowan, "A sector-based approach for localization of multiple speakers with microphone arrays," in *Proceedings of the Workshop on Statistical and Perceptual Audio Processing (SAPA '04)*, Jeju, Republic of Korea, 2004.

[38] D. Vijayasenan, F. Valente, and H. Bourlard, "An information theoretic approach to speaker diarization of meeting data," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 17, no. 7, pp. 1382–1393, 2009.

[39] A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki, "The DET curve in assessment of detection task performance," in *Proceedings of the European Conference on*

Speech Communication and Technology (Eurospeech '97), vol. 4, pp. 1895–1898, Rhodes, Greece, 1997.

[40] EBU Technical Recommendation, "MUSHRA-EBU method for subjective listening tests of intermediate audio quality," Doc. B/AIM022, 1999.