Speech Enhancement and Recognition in Meetings With an Audio–Visual Sensor Array

Hari Krishna Maganti, *Student Member, IEEE*, Daniel Gatica-Perez, *Member, IEEE*, and Iain McCowan, *Member, IEEE*

***Abstract*—This paper addresses the problem of distant speech acquisition in multiparty meetings, using multiple microphones and cameras. Microphone array beamforming techniques present a potential alternative to close-talking microphones by providing speech enhancement through spatial filtering. Beamforming techniques, however, rely on knowledge of the speaker location. In this paper, we present an integrated approach, in which an audio–visual multiperson tracker is used to track active speakers with high accuracy. Speech enhancement is then achieved using microphone array beamforming followed by a novel postfiltering stage. Finally, speech recognition is performed to evaluate the quality of the enhanced speech signal. The approach is evaluated on data recorded in a real meeting room for stationary speaker, moving speaker, and overlapping speech scenarios. The results show that the speech enhancement and recognition performance achieved using our approach are significantly better than a single table-top microphone and are comparable to a lapel microphone for some of the scenarios. The results also indicate that the audio–visual-based system performs significantly better than audio-only system, both in terms of enhancement and recognition. This reveals that the accurate speaker tracking provided by the audio–visual sensor array proved beneficial to improve the recognition performance in a microphone array-based speech recognition system.**

***Index Terms*—Audio–visual fusion, microphone array processing, multiobject tracking, speech enhancement, speech recognition.**

I. Introduction

With the advent of ubiquitous computing, a significant trend in human–computer interaction is the use of a range of multimodal sensors and processing technologies to observe the user's environment. These allow users to communicate and interact naturally, both with computers and with other users. Example applications include advanced computing environments [1], instrumented meeting rooms [44], [54], and seminar halls [10] facilitating remote collaboration. The current article examines the use of multimodal sensor arrays in the context of instrumented meeting rooms. Meetings consist of natural, complex interaction between multiple participants, and so automatic analysis of meetings is a rich research area, which has been studied actively as a motivating application for a range of multidisciplinary research [25], [44], [47], [54].

Speech is the predominant communication mode in meetings. Speech acquisition, processing, and recognition in meetings are complex tasks, due to the nonideal acoustic conditions (e.g., reverberation, noise from presentation devices, and computers usually present in meeting rooms) as well as the unconstrained nature of group conversation in which speakers often move around and talk concurrently. A key goal of speech processing and recognition systems in meetings is the acquisition of high-quality speech without constraining users with tethered or close-talking microphones. Microphone arrays provide a means of achieving this through the use of beamforming techniques.

A key component of any practical microphone array speech acquisition system is the robust localization and tracking of speakers. Tracking speakers solely based on audio is a difficult task due to a number of factors: human speech is an intermittent signal, speech contains significant energy in the low-frequency range, where spatial discrimination is imprecise, and location estimates are adversely affected by noise and room reverberations. For these reasons, a body of recent work has investigated an audio–visual approach to speaker tracking in conversational settings such as videoconferences [28] and meetings [9]. To date, speaker tracking research has been largely decoupled from microphone array speech recognition research. With the increasing maturity of approaches, it is timely to properly investigate the combination of tracking and recognition systems in real environments, and to validate the potential advantages that the use of multimodal sensors can bring for the enhancement and recognition tasks.

The present work investigates an integrated system for hands-free speech recognition in meetings based on an audio–visual sensor array, including a multimodal approach for multiperson tracking, and speech enhancement and recognition modules. Audio is captured using a circular, table-top array of eight microphones, and visual information is captured from three different camera views. Both audio and visual information are used to track the location of all active speakers in the meeting room. Speech enhancement is then achieved using microphone array beamforming followed by a novel postfiltering stage. The enhanced speech is finally input into a standard hidden Markov model (HMM) recognizer system to

Manuscript received April 18, 2006; revised February 20, 2007. This work was supported by the European Projects Augmented Multi-Party Interaction (AMI, EU-IST project FP6-506811, pub. AMI-239), Detection and Identifications of Rare Audio-Visual Cues (DIRAC, EU-IST project FP6-003758), and by the Swiss National Center for Comptetence in Research (NCCR) on Interactive Multimodal Information Management (IM2). The associate editor coordinating the review of this manuscript and approving it for publication was Dr. George Tzanetakis.

H. K. Maganti is with the Institute of Neural Information Processing University of Ulm, D-89069 Ulm, Germany.

D. Gatica-Perez is with the IDIAP Research Institute and Ecole Polytechnique Federale de Lausanne (EPFL), CH-1920 Martigny, Switzerland.

I. McCowan is with the CSIRO eHealth Research Centre and Speech and Audio Research Laboratory, Queensland University of Technology, QLD 4000 Brisbane, Australia.

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

Digital Object Identifier 10.1109/TASL.2007.906197

evaluate the quality of the speech signal. Experiments consider three scenarios common in real meetings: a single seated active speaker, a moving active speaker, and overlapping speech from concurrent speakers. To investigate in detail the subsequent effects of tracking on speech enhancement and recognition, the study has been confined to the specific cases of one and two speakers around a meeting table. The speech recognition performance achieved using our approach is compared to that achieved using headset microphones, lapel microphones, and a single table-top microphone. To quantify the advantages of a multimodal approach to tracking, results are also presented using a comparable audio-only system. The results show that the audio–visual tracking-based microphone array speech enhancement and recognition system performs significantly better than single table-top microphone and comparable to lapel microphone for all the scenarios. The results also indicate that the audio–visual-based system performs significantly better than audio-only system in terms of signal-to-noise ratio enhancement (SNRE) and word error rate (WER). This demonstrates that the accurate speaker tracking provided by the audio–visual sensor array improves speech enhancement, in turn resulting in improved speech recognition performance.

This paper is organized as follows: Section II discusses the related work. Section III gives an overview of the proposed approach. Section IV describes the sensor array configuration and intermodality calibration issues. Section V details the audio–visual person tracking technique. Section VI presents the speech enhancement module, while speech recognition is described in Section VII. Section VIII presents the data, the experiments, and their discussion, and finally conclusions are given in Section IX.

II. Related Work

Most state-of-the-art speech processing systems rely on close-talking microphones for speech acquisition, as they naturally provide the best performance. However, in multiparty conversational settings like meetings, this mode of acquisition is often not suitable, as it is intrusive and constrains the natural behavior of a speaker. For such scenarios, microphone arrays present a potential solution by offering distant, hands-free, and high-quality speech acquisition through beamforming techniques [52].

Beamforming consists of filtering and discriminating active speech sources from various noise sources based on location. The simplest beamforming technique is delay-sum, in which a delay filter is applied to each microphone channel before summing them to give a single enhanced output. A more sophisticated filter-sum beamformer that has shown good performance in speech processing applications is superdirective beamforming, in which filters are calculated to maximize the array gain for the look direction [13]. The post filtering of the beamformer output significantly improves desired signal enhancement by reducing background noise [38]. Microphone array speech recognition, i.e, the integration of a beamformer with automatic speech recognition for meeting rooms has been investigated in [45]. In the same context, in National Institute of Standards and Technology (NIST) meeting recognition evaluations, techniques were evaluated to recognize the speech from multiple distant microphones, with systems required to handle varying numbers of microphones, unknown microphone placements, and an unknown number of speakers [47].

The localization and tracking of multiple active speakers are crucial for optimal performance of microphone-array-based speech acquisition systems. Many computer vision systems [8], [14] have been studied to detect and track people, but are affected by occlusion and illumination effects. Acoustic source localization algorithms can operate in different lighting conditions and localize in spite of visual occlusions. Most acoustic source localization algorithms are based on the time-difference of arrival (TDOA) approach, which estimate the time delay of sound signals between the microphones in an array. The generalized cross-correlation phase transform (GCC-PHAT) method [32] is based on estimating the maximum GCC between the delayed signals and is robust to reverberations. The steered response power (SRP) method [33] is based on summing the delayed signals to estimate the power of output signal and is robust to background noise. The advantages of both the methods, i.e, robustness to reverberations and background noise are combined in the SRP-PHAT method [15]. To enhance the accuracy of TDOA estimates and handle multispeaker cases, Kalman filter smoothing was studied in [51] and combination of TDOA with particle filter approach has been investigated in [55]. However, due to the discreteness and vulnerability to noise sources and strong room reverberations, tracking based exclusively on audio estimates is an arduous task. To account for these limitations, multimodal approaches combining acoustic and visual processing have been pursued recently for single-speaker [2], [4], [19], [53], [59] and multispeaker [7], [9], [28] tracking. As demonstrated by the tasks defined in the recent Classifications of Events, Actions, and Relations (CLEAR) 2006 evaluation workshop, multimodal approaches constitute a very active research topic in the context of seminar and conference rooms to track presenters, or other active speakers [6], [29], [46]. In [29], a 3-D tracking with stand-alone video and audio trackers was combined using a Kalman filter. In [46], it was demonstrated that the audio–visual combination yields significantly greater accuracy than either of the modalities. The proposed algorithm was based on a particle filter approach to integrate acoustic source localization, person detection, and foreground segmentation using multiple cameras and multiple pairs of microphones. The goal of fusion is to make use of complementary advantages: initialization and recovery from failures can be addressed with audio, and precise object localization with visual processing [20], [53].

Being major research topics, speaker tracking and microphone array speech recognition have recently reached levels of performance where they can start being integrated and deployed in real environments. Recently, Asano *et al.* presented a framework where a Bayesian network is used to detect speech events by the fusion of sound localization from a small microphone array and vision tracking based on background subtraction from two cameras [2]. The detected speech event information was used to vary beamformer filters for enhancement, and also to separate desired speech segments from noise in the enhanced speech, which was then used as input to the speech recognizer. In other recent work, particle filter data fusion with audio from multiple large microphone arrays and video from multiple calibrated cameras was used in the context of seminar rooms [39]. The audio features were based on time delay of arrival estimation. For the video features, dynamic foreground segmentation based on adaptive background modeling as a primary feature along with foreground detectors were used. The system assumes that the lecturer is the person standing and moving while the members of the audience are sitting and moving less, and that there is essentially one main speaker (the lecturer). As we describe in the remainder of this paper, our work substantially differs from previous works in the specific algorithms used for localization, tracking, and speech enhancement. Our paper is focused on robust speech acquisition in meetings and specifically has two advantages over [2] and [39]. First, our tracking module can track multiple speakers irrespective of the state of the speakers, e.g., seated, standing, fixed, or moving. Second, in the enhancement module, the beamformer is followed by a postfilter which helps in broadband noise reduction of the array, leading to better performance in speech recognition. Finally, our sensor setup aims at dealing with small group discussions and relies on a small microphone array, unlike [39] which relies on large arrays. For the appraisal of the tracking effects on speech enhancement and recognition, our experiments were limited to the cases of one and two speakers around a table in a meeting room (other recent studies, including works in the CLEAR evaluation workshop, have handled other scenarios, like presenters in seminars). A preliminary version of our work was presented in [42].

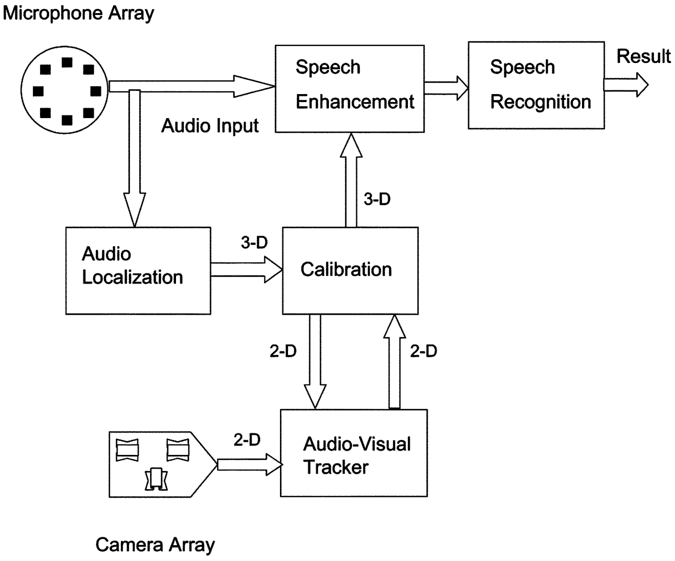

Fig. 1. System block diagram. The microphone array provides audio inputs to the speech enhancement and audio localization modules. Three-dimensional localization estimates are generated by the audio localization module, which are mapped onto the corresponding 2-D image plane by the calibration module. The audio–visual tracker processes this 2-D information along with the visual information from the camera array to track the active speakers. The 3-D estimates are reconstructed by the calibration module from two camera views, which are then input to the speech enhancement module. The enhanced speech from the speech enhancement module, which is composed of a beamformer followed by a postfilter, is used as input to the speech recognition module.

III. Overview of Our Approach

A schematic description of our approach is shown in Fig. 1. The goal of the blocks on the bottom left part of the figure (Audio Localization, Calibration, and Audio–Visual Tracker) is to accurately estimate, at each time-step, the 3-D locations of each of the people present in a meeting, $\hat{Z}_t = (\hat{Z}_{i,t})$, $i \in \mathcal{I}_t$, where $\mathcal{I}_t$ is the set of person identifiers, $\hat{Z}_{i,t}$ denotes the location for person i, and $m_t = |\mathcal{I}_t|$ denotes the number of people in the scene. The estimation of location is done with a multimodal approach, where the information captured by the audio–visual sensors is processed to exploit the complementarity of the two modalities. Human speech is discontinuous in nature. This represents a fundamental challenge for tracking location based solely on audio, as silence periods imply, in practice, lack of observations: people might silently change their location in a meeting (e.g., moving from a seat to the white board) without providing any audio cues that allow for either tracking in the silent periods or reidentification. In contrast, video information is continuous, and person location can in principle be continuously inferred through visual tracking. On the other hand, audio cues are useful, whenever available, to robustly reinitialize a tracker, and to keep a tracker in place when visual clutter is high.

Our approach uses data captured by a fully calibrated audio–visual sensor array consisting of three cameras and a small microphone array, which covers the meeting workspace with pair-wise overlapping views, so that each area of the workspace of interest is viewed by two cameras. The sensor array configuration and calibration are further discussed in Section IV. In our methodology, the 2-D location of each person visible in each camera plane is continuously estimated using a Bayesian multiperson state-space approach. The multiperson state configurations in camera plane k are defined as $X_t^k = \left(X_{i,t}^k\right)$, $i \in \mathcal{I}_t$, where $\mathcal{I}_t$ is the set of person identifiers mentioned above, and $X_{i,t}^k$ denotes the configuration of person i. For audio–visual observations $Y_t^k = \left(Y_t^{k,a}, Y_t^{k,v}\right)$, where the vector components $Y_t^{k,a}$ and $Y_t^{k,v}$ denote the audio and visual observations, respectively, the filtering distribution of states given observations $p\left(X_t^k | Y_{1:t}^k\right)$ is recursively approximated using a Markov Chain Monte Carlo (MCMC) particle filter [21].

This algorithm is described in Section V. For this, a set of 3-D audio observations $\left\{Y_t^{a3}\right\}$ is derived at each time-step using a robust source localization algorithm based on the SRP-PHAT measure [34]. Using the sensor calibration method described in Section IV, these observations are mapped onto the two corresponding camera image planes by a mapping function $f_\Phi : \mathbb{R}^3 \rightarrow (\{0,1,2\} \times \mathbb{R}^2)^2$, where Φ indicates the camera calibration parameters, which associates a 3-D position with a 6-D vector containing the camera index k_t^j and the 2-D image position $Y_t^{k_t^j,a}$ for the corresponding pair of camera planes $j \in \{1,2\}$. Visual observations are extracted from the corresponding image planes. Finally, for each person i, the locations estimated by the trackers, $\hat{X}_{i,t}^{k_t^1}$, $\hat{X}_{i,t}^{k_t^2}$ for the corresponding camera pair, k_t^1 and k_t^2, are merged. The corresponding 3-D location estimate is obtained using the inverse mapping $\widehat{Z_{i,t}} = f_\Phi^{-1}\left(k_t^1, \hat{X}_{i,t}^{k_t^1}, k_t^2, \hat{X}_{i,t}^{k_t^2}\right)$.

The 3-D estimated locations for each person are integrated with the beamformer as described in Section VI. At each time-step, for which the distance between the tracked speaker location and the beamformer's focus location exceeds a small value, the beamformer channel filters are recalculated. For further speech signal enhancement, the beamformer is followed by a postfiltering stage. After speech enhancement, speech recognition is performed on the enhanced signal. This is discussed in Section VII. In summary, a baseline speech recognition system is first trained using the headset microphone data from the original Wall Street Journal corpus [49]. A number of adaptation techniques, including maximum-likelihood linear regression (MLLR) and maximum *a posteriori* (MAP), are used to compensate for the channel mismatch between the training and test conditions. Finally, to fully compare the effects of audio versus audio–visual estimation of location in speech enhancement and recognition, the audio-only location estimates directly computed from the speaker localization module in Fig. 1 are also fed into the enhancement and recognition blocks of our approach.

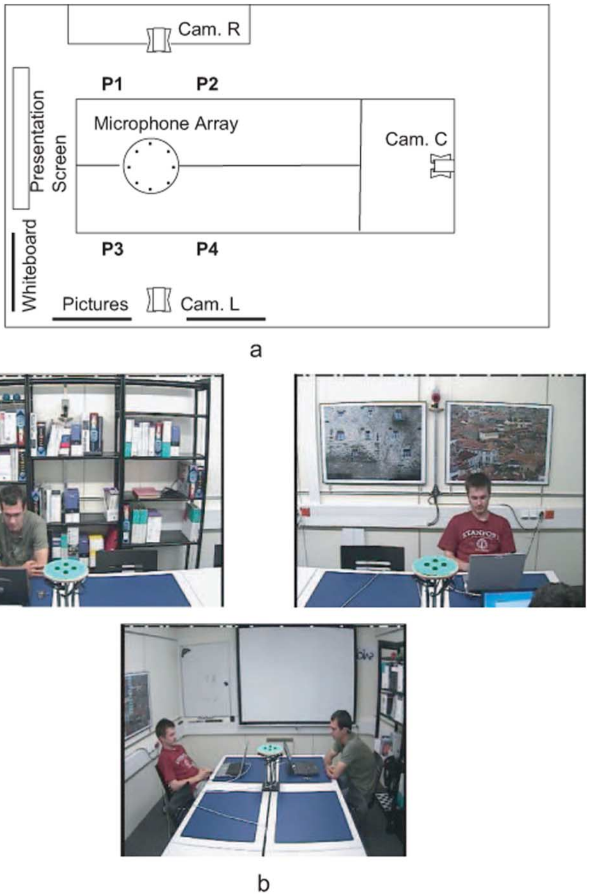

Fig. 2. (a) Schematic diagram of the meeting room. Cam. C, L, and R denote the center, left, and right cameras, respectively (referred to as cameras 0,1, and 2 in Section III). P1, P2, P3, and P4 indicate the typical speaker positions. (b) Left, right, and center sample images. The meeting room contains visual clutter due to bookshelves and skin-colored posters. Audio clutter is caused from the laptops and other computers in the room. Speakers act naturally with no constraints on speaking styles or accents.

IV. Audio–Visual Sensor Array

A. Sensor Configuration

All the data used for experiments are recorded in a moderately reverberant multisensor meeting room. The meeting room is a 8.2 m × 3.6 m × 2.4 m containing a 4.8 m × 1.2 m rectangular table at one end [45]. Fig. 2(a) shows the room layout, the position of the microphone array and the video cameras, and typical speaker positions in the room. The sample images of the three views from the meeting room are as shown in Fig. 2(b). The audio sensors are configured as an eight-element, circular equi-spaced microphone array centered on the table, with diameter of 20 cm, and composed of high-quality miniature electret microphones. Additionally, lapel and headset microphones are used for each speaker. The video sensors include three wide-angle cameras (center, left, and right) giving a complete view of the room. Two cameras on opposite walls record frontal views of participants, including the table and workspace area, and have nonoverlapping fields-of-view (FOVs). A third wide-view camera looks over the top of the participants towards the white-board and projector screen. The meeting room allows capture of fully synchronized audio and video data.

B. Sensor Calibration

To relate points in the 3-D camera reference with 2-D image points, we calibrate the three cameras (center, left, and right) of the meeting room to a single 3-D external reference using a standard camera calibration procedure [58]. This method, with a given number of image planes represented by a checkerboard at various orientations, estimates the different camera parameters which define an affine transformation relating the camera reference and the 3-D external reference. The microphone array has its own external reference, so in order to map a 3-D point in the microphone array reference to an image point, we also define a transformation for basis change between the microphone array reference and the 3-D external reference. Finally, to complete the audio–video mapping, we find the correspondence between image points and 3-D microphone array points. From stereovision, the 3-D reconstruction of a point can be done with the image coordinates of the same point in two different camera views. Each point in each camera view defines a ray in 3-D space. Optimization methods are used to find the intersection of the two rays, which corresponds to the reconstructed 3-D point [26]. This last step is used to map the output of the audio–visual tracker (i.e., the speaker location in the image planes) back to 3-D points, as input to the speech enhancement module.

V. Person Tracking

To jointly track multiple people in each image plane, we use the probabilistic multimodal multispeaker tracking method proposed in [21], consisting of a dynamic Bayesian network in which approximate inference is performed by an MCMC particle filter [18], [36], [30]. In the rest of the section, we describe the most important details of the method in [21] for purposes of completeness. Furthermore, to facilitate reading, the notation is simplified with respect to Section III by dropping the camera index symbol, so multiperson configurations $X_t^k = \left(X_{i,t}^k\right)$ are denoted by $X_t = (X_{i,t})$, observations Y_t^k by Y_t, etc.

Given a set of audio–visual observations $Y_{1:t}$, and a multi-object mixed state-space X_t, defined by continuous geometric transformations (e.g., motion) and discrete indices (e.g., of the speaking status) for multiple people, the filtering distribution $p(X_t|Y_{1:t})$ can be recursively computed using Bayes' rule by

$$p(X_t|Y_{1:t}) \propto p(Y_t|X_t) \cdot \int_{X_{t-1}} p(X_t|X_{t-1})p(X_{t-1}|Y_{1:t-1})dX_{t-1}, \quad (1)$$

where $p(X_t|X_{t-1})$ denotes the multiperson dynamical model, and $p(Y_t|X_t)$ denotes the multiperson observation model. A particle filter recursively approximates the filtering distribution $p(X_t|Y_{1:t})$ by a weighted set of N_s particles $\left\{X_t^{(n)}, w_t^{(n)}\right\}_{n=1}^{N_s}$, using the particle set at the previous time-step, $\left\{X_{t-1}^{(n)}, w_{t-1}^{(n)}\right\}$, and the new observations

$$p(X_t|Y_{1:t}) \approx \mathcal{Z}^{-1}p\left(Y_t|X_t\right)\sum_{n=1}^{N_s} w_{t-1}^{(n)}p\left(X_t|X_{t-1}^{(n)}\right) \quad (2)$$

where $\mathcal{Z}$ denotes a normalization constant. In our paper, the multiperson state-space is composed of mixed state-spaces defined for each person's configuration $X_{i,t}$ that include 1) a continuous vector of transformations—including 2-D translation and scaling—of a person's head template—an elliptical silhouette—in the image plane, and 2) a discrete binary variable modeling the person speaking activity $X_{i,t} = (x_{i,t}, l_{i,t})$. As can be seen from 2), the three key elements of the approach are the dynamical model, the observation likelihood model, and the sampling mechanism which are discussed in the following three subsections.

A. Dynamical Model

The dynamical model includes both independent single-person dynamics and pairwise interactions. A pairwise Markov random field (MRF) prior constrains the dynamics of each person based on the state of the others [30]. The MRF is defined on an undirected graph, where objects define the vertices, and links exist between object pairs at each time-step. With these definitions, the dynamical model is given by

$$p(X_t|X_{t-1}) \propto \left(\prod_{i\in\mathcal{I}_t} p(X_{i,t}|X_{i,t-1})\right) \times \left(\prod_{(i,j)\in\mathcal{C}} \phi(X_{i,t}, X_{j,t})\right) \quad (3)$$

where $p(X_{i,t}|X_{i,t-1})$ denote the single-object dynamics, and the prior is the product of potentials $\phi(X_{i,t}, X_{j,t})$ over the set $\mathcal{C}$ of pairs of connected nodes in the graph. Equation (2) can then be expressed as

$$p(X_t|Y_{1:t}) \approx \mathcal{Z}^{-1}p(Y_t|X_t)\left(\prod_{(i,j)\in\mathcal{C}} \phi(X_{i,t}, X_{j,t})\right) \times \left(\sum_n w_{t-1}^{(n)} \prod_{i\in\mathcal{I}_t} p(X_{i,t}|X_{i,t-1}^{(n)})\right). \quad (4)$$

The dynamical model for each object is defined as

$$p(X_{i,t}|X_{i,t-1}) = p(x_{i,t}|x_{i,t-1})p(l_{i,t}|l_{i,t-1})$$

where the continuous distribution $p(x_{i,t}|x_{i,t-1})$ is a second-order autoregressive model [27], and $p(l_{i,t}|l_{i,t-1})$ is a 2×2 transition probability matrix (TPM).

The possibility of associating two configurations to one single object when people occlude each other momentarily is handled by the interaction model, which penalizes large overlaps between objects [30]. For any object pair $X_{i,t}$ and $X_{j,t}$ with spatial supports $\mathcal{S}_{i,t}$ and $\mathcal{S}_{j,t}$, respectively, the pairwise overlap measures are the typical precision $(\nu(\mathcal{S}_{i,t}, \mathcal{S}_{j,t}))$ and recall $(\rho(\mathcal{S}_{i,t}, \mathcal{S}_{j,t}))$. The pairwise potentials in the MRF $\phi(X_{i,t}, X_{j,t})$ are defined by an exponential distribution over precision/recall features.

B. Observation Model

The observation model is derived from both audio and video. Audio observations are derived from a speaker localization algorithm, while visual observations are based on shape and spatial structure of human heads. The observations are defined as $Y_t = (Y_t^a, Y_t^v)$, where $Y_t^v = \left(Y_t^{sh}, Y_t^{st}\right)$, and the superindices stand for *a*udio, *v*ideo *sh*ape, and spatial *st*ructure, respectively. The observations are assumed to be conditionally independent given the single-object states

$$p(Y_t|X_t) = \prod_{i\in\mathcal{I}_t} p(Y_{i,t}^a|X_{i,t})p(Y_{i,t}^{sh}|X_{i,t})p(Y_{i,t}^{st}|X_{i,t}). \quad (5)$$

A sector-based source localization algorithm is used to generate the audio observations, in which candidate 3-D locations of the participants are computed when people speak. The work in [34] proposed a simple source localization algorithm, which utilizes low computational resources and is suitable for reverberant environments, based on the steered response power—phase transform (SRP-PHAT) technique [16]. In this approach, a fixed grid of points is built by selecting points on a set of concentric spheres centered on the microphone array. Given that the sampling rate for audio is higher than the one for video, multiple audio localization estimates (between zero and three) are available at each video frame. We then use the sensor calibration procedure in the previous section to project the 3-D audio estimates on the corresponding 2-D image planes. Finally, the audio observation likelihood $p\left(Y_{i,t}^a|X_{i,t}\right)$ is defined as a switching distribution (depending on the predicted value of the binary speaking activity variable $l_{i,t}$) over the Euclidean distance between the projected 2-D audio localization estimates and the translation components of the candidate configurations $x_{i,t}$. The switching observation model satisfies the notion that, if a person is predicted to be speaking, an audio-estimate should exist and be near such person, while if a person is predicted to be silent, no audio estimate should exist or be nearby.

The visual observations are based on shape and spatial structure of human heads. These two visual cues complement each other, as the first one is edge-oriented while the second one is region-oriented. The shape observation model is derived from a classic model in which edge features are computed over a number of perpendicular lines to a proposed elliptical head configuration [27]. The shape likelihood $p\left(Y_{i,t}^{sh}|X_{i,t}\right)$ is defined over these observations. The spatial structure observations are based on a part-based parametric representation of the overlap between skin-color blobs and head configurations. Skin-color blobs are first extracted at each frame according to a standard procedure described in [20], based on a Gaussian mixture model (GMM) representation of skin color. Then, precision/recall overlap features, computed between the spatial supports of skin-color blobs and the candidate configurations, represented by a part-based head model, are extracted. This feature representation aims at characterizing the specific distribution of skin-color pixels in the various parts of a person's head. The spatial structure likelihood $p\left(Y_{i,t}^{st}|X_{i,t}\right)$ is a GMM defined over the precision/recall features. Training data for the skin-color model and the spatial structure model is collected from people participating in meetings in the room described in Section IV.

C. Sampling Mechanism

The approximation of (4) in the high-dimensional space defined by multiple people is done with MCMC techniques, more specifically designing a Metropolis–Hastings sampler at each time step in order to efficiently place samples as close as possible to regions of high likelihood [30]. For this purpose, we define a proposal distribution in which the configuration of one single object is modified at each step of the Markov chain, and each move in the chain is accepted or rejected based on the evaluation of the so-called acceptance ratio in the Metropolis–Hastings algorithm. This proposal distribution results in a computationally efficient acceptance ratio calculation [21]. After discarding an initial burn-in set of samples, the generated MCMC samples will approximate the target filtering distribution [36]. A detailed description of the algorithm can be found in [22].

At each time-step, the output of the multiperson tracker is represented by the mean estimates for each person. From here, the 2-D locations of each person's head center for the specific camera pair where such person appears, which correspond to the translation components of the mean configuration in each camera and are denoted by $\hat{X}_{i,t}^{k_1^t}$, $\hat{X}_{i,t}^{k_2^t}$, can be extracted and triangulated as described in Section IV-B to obtain the corresponding 3-D locations $\hat{Z}_{i,t}$. These 3-D points are finally used as input to the speech enhancement module, as described in Section VI.

VI. Speech Enhancement

The microphone array speech enhancement system includes a filter-sum beamformer, followed by a postfiltering stage, as shown in Fig. 3. The superdirective technique was used to calculate the channel filters maximizing the array gain, while maintaining a minimum constraint on the white noise gain. This technique is fully described in [41]. The optimal filters are calculated as

$$\mathbf{w}_{\text{opt}} = \frac{\mathbf{\Gamma}^{-1}\mathbf{d}}{\mathbf{d}^H\mathbf{\Gamma}^{-1}\mathbf{d}} \tag{6}$$

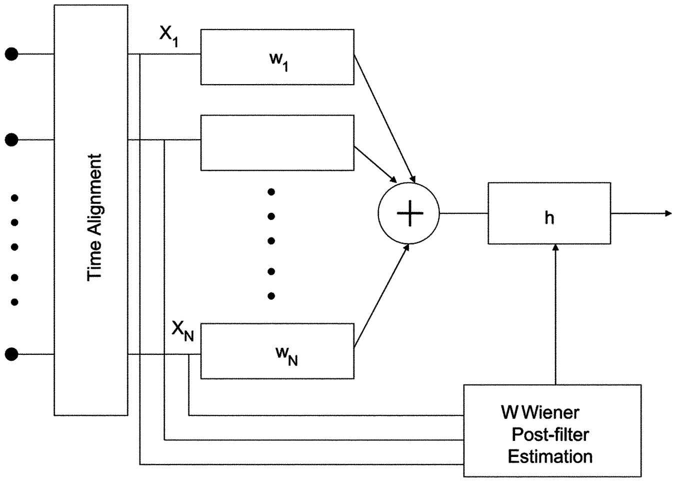

Fig. 3. Speech enhancement module with filter-sum beamformer followed by a postfilter.

where $\mathbf{w}_{\text{opt}}$ is the vector of N optimal filter coefficients

$$\mathbf{w}_{\text{opt}}(f) = [w_1(f) \quad w_2(f) \quad \dots \quad w_N(f)]^T \tag{7}$$

where f denotes frequency, and $\mathbf{d}$ is the propagation vector between the source and each microphone

$$\mathbf{d}(f) = [\alpha_1 e^{-2\pi f \delta_1} \quad \alpha_2 e^{-2\pi f \delta_2} \quad \dots \quad \alpha_N e^{-2\pi f \delta_N}]^T \tag{8}$$

$\mathbf{\Gamma}$ is the noise coherence matrix (assumed diffuse), and α_n, δ_n are the channel scaling factors and delays due to the propagation distance.

As an illustration of the expected directivity from such a superdirective beamformer, Fig. 4 shows the polar directivity pattern at several frequencies for the array used, calculated at a distance of 1 m from the array center. The geometry gives reasonable discrimination between speakers separated by at least 45°, making it suitable for small group meetings of up to eight participants (assuming a relatively uniform angular distribution of participants). For the experiments in this paper, we integrated the tracker output with the beamformer in a straightforward manner. Any time the distance between the tracked speaker location and the beamformer's focus location exceeded 2 cm, the beamformer channel filters were recalculated.

A. Postfilter for Overlapping Speech

The use of a postfilter following the beamformer has been shown to improve the broadband noise reduction of the array [38], and lead to better performance in speech recognition applications [45]. Much of this previous work has been based on the use of the (time-delayed) microphone auto- and cross- spectral densities to estimate a Wiener transfer function. While this approach has shown good performance in a number of applications, its formulation is based on the assumption of low correlation between the noise on different microphones. This assumption clearly does not hold when the predominant "noise" source is coherent, such as overlapping speech. In the following, we propose a new postfilter better suited for this case.

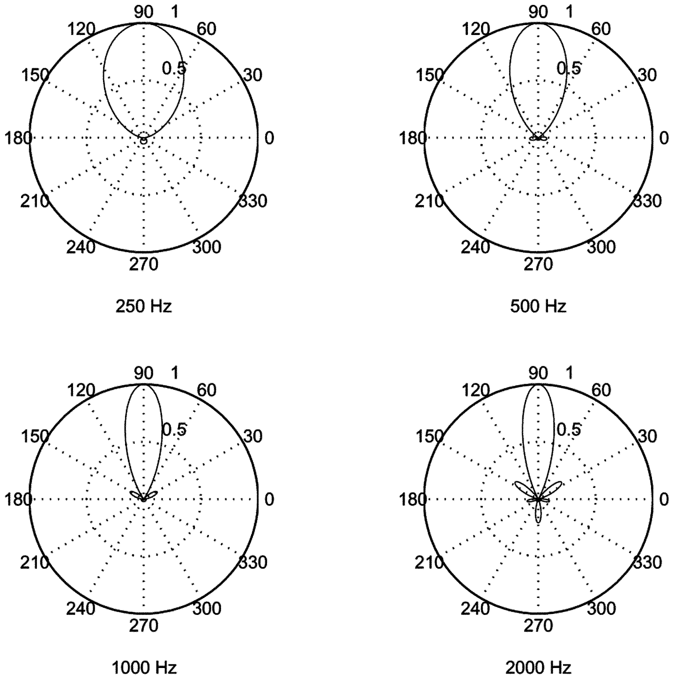

Fig. 4. Horizontal polar plot of the near-field directivity pattern ($r = 1$ m) of the superdirective beamformer for an eight-element circular array of radius 10 cm.

Assume that we have S beamformers concurrently tracking S different people within a room, with frequency-domain outputs $b_s(f)$, $s = 1 : S$. We further assume that in each $b_s(f)$, the energy of speech from person s (when active) is higher than the energy level of all other people. It has been observed (see [50] for a discussion) that the log spectrum of the additive combination of two speech signals can be well approximated by taking the maximum of the two individual spectra in each frequency bin, at each time. This is essentially due to the sparse and varying nature of speech energy across frequency and time, which makes it highly unlikely that two concurrent speech signals will carry significant energy in the same frequency bin at the same time. This property was exploited in [50] to develop a single-channel speaker separation system.

We apply the above property over the S frequency-domain beamformer outputs to calculate S simple masking postfilters $h_s(f)$

$$h_s = \begin{cases} 1, & \text{if } s = \underset{s'}{\arg\max} \; |b_s(f)|^2, \; s' = 1 : S \\ 0, & \text{otherwise.} \end{cases} \tag{9}$$

Each post-filter is then applied to the corresponding beamformer output to give the final enhanced output of the person s as $u_{i,s}(f) = h_{i,s}(f) b_{i,s}(f)$, where i is the spectrogram frame index. Note that when only one person is actively speaking, other beamformers essentially provide an estimate of the background noise level, and therefore the postfilter would function to reduce the background noise. To achieve such an effect in the single-speaker experimental scenarios, a second beamformer is oriented to the opposite side of the table for use in the above postfilter. This has a benefit of low computational cost compared to other formulations such as those based on the calculation of channel auto- and cross-spectral densities [57].

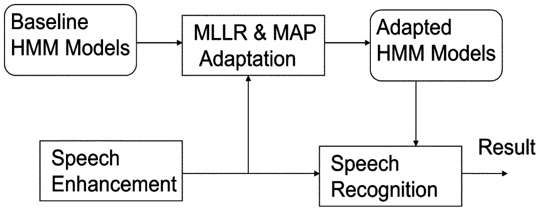

Fig. 5. Speech recognition adaptation. The baseline HMM models are adapted using MLLR and MAP techniques. The acoustics of the enhanced speech signal from speech enhancement block are adjusted to improve the speech recognition performance.

VII. Speech Recognition

With the ultimate goal of automatic speech recognition, speech recognition tests are performed for the stationary, moving speaker, and overlapping speech scenarios. This is also important to quantify the distortion to the desired speech signal. For the baseline, a full HTK-based recognition system, trained on the original Wall Street Journal database (WSJCAM0) is used [49]. The training set consists of 53 male and 39 female speakers, all with British English accents. The system consists of approximately 11 000 tied-state triphones with three emitting states per triphone and six mixture components per state. 52-element feature vectors were used, comprising of 13 Mel cepstral frequency coefficients (MFCCs) (including the 0th cepstral coefficient) with their first-, second-, and third-order derivatives. Cepstral mean normalization is performed on all the channels. The dictionary used is generated from that developed for the Augmented Multiparty Interaction (AMI) project and used in the evaluations of the National Institute of Standards and Technology Rich Transcription (NIST RT05S) system [25], and the language model is the standard MIT-Lincoln Labs 20k Wall Street Journal (WSJ) trigram language model. The baseline system with no adaptation gives 20.44% WER on the si_dt20a task (20 000 word), which roughly corresponds to the results reported in the SQALE evaluation using the WSJCAM0 database [56].

To reduce the channel mismatch between the training and test conditions, the baseline HMM models are adapted using maximum-likelihood linear regression (MLLR) [35] and MAP [23] adaptation as shown in Fig. 5. Adaptation data was matched to the testing condition (that is, headset data was used to adapt models for headset recognition, lapel data was used to adapt for lapel recognition, etc.).

VIII. Experiments and Results

Sections VIII-A–D describe the database specification, followed by tracking, speech enhancement, and speech recognition results. The results, along with additional meeting room data results for a single speaker switching seats, and for overlap speech from two side-by-side simultaneous speakers can be viewed at the companion website http://www.idiap.ch/~hakri/avsensorarray/avdemos.htm.

A. Database Specification

All the experiments are conducted on a subset of the Multi-Channel Wall Street Journal Audio-Visual (MC-WSJ-AV) corpus. The specification and structure of the

TABLE I
DATA DESCRIPTION

Scenario	No. of sentences	Total time (min.)	No. of speakers
Stationary	160	22	6
Moving	78	12	6
Overlap	70	11	4

full corpus are detailed in [37]. We used a part of *single-speaker stationary*, *single-speaker moving*, and *stationary overlapping speech* data from the 20k WSJ task. In the single-speaker stationary case, the speaker reads out sentences from different positions within the meeting room. In the single-speaker moving scenario, the speaker is moving between different positions while reading the sentences. Finally, in the overlapping speech case, two speakers simultaneously read sentences from different positions within the room. Most of the data comprised of nonnative English speakers with different speaking styles and accents. The data is divided into development (DEV) and evaluation (EVAL) sets with no common speakers across sets. Table I describes the data used for the experiments.

B. Tracking Experiments

The multiperson tracking algorithm was applied to the data set described in the previous section, for each of the three scenarios (stationary single-person, moving single-person, and two-person overlap). In the tracker, all models that require a learning phase (e.g., the spatial structure head model), and all parameters that are manually set (e.g., the dynamic model parameters), were learned or set using a separate data set, originally described in [21], and kept fixed for all experiments. Regarding the number of particles, experiments were done for 500, 800, and 500 particles for the stationary, moving, and overlap cases, respectively. In all cases, 30% of the particles were discarded during the burn-in period of the MCMC sampler, and the rest were kept for representing the filtering distribution at each time-step. It is important to notice that the number of particles was not tuned but simply set to a sensible fixed value, following the choices made in [21]. While the system could have performed adequately with less particles, the dependence on the number of particles was not investigated here. All reported results are computed from a single run of the tracker.

The accuracy of tracking was objectively evaluated by the following procedure. The 3-D Euclidean distance between a ground truth location of the speakers mouth represented by (x, y, z) and the automatically estimated location $(\hat{x}, \hat{y}, \hat{z})$ was used as performance measure. For N frames, this was computed as

$$\mathrm{Ed} = \frac{1}{N}\sum_{n=1}^{N}\sqrt{(x_n - \hat{x}_n)^2 + (y_n - \hat{y}_n)^2\,(z_n - \hat{z}_n)^2}. \quad (10)$$

The frame-based ground truth was generated as follows. First, the 2-D point mouth position of each speaker was manually annotated in each camera plane. Then, each pair of 2-D points was reconstructed into a 3-D point using the inverse mapping. The ground truth was produced at a rate of 1 frame/s every 25 video frames. The 3-D Euclidean distance is averaged over all frames in the data set.

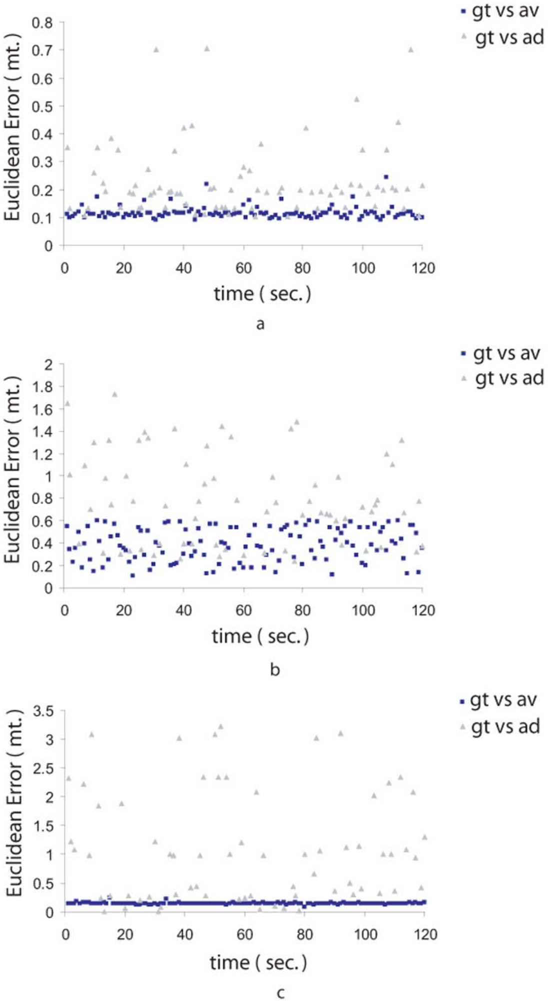

Fig. 6. Tracking results for (a) stationary, (b) moving speaker, and (c) overlapping speech, for 120 s of video for each scenario. "gt versus av" and "gt versus ad" represent ground truth versus audio–visual tracker, and ground truth versus output of the audio-only localization algorithm, respectively. For audio-only, the "average" error is computed (see text for details). Audio estimates are discontinuous and available around 60% of the times. The audio–visual estimates are continuous and more stable.

The results are presented in Table II, Fig. 6, and on the companion website. Table II summarizes the overall results, Fig. 6 illustrates typical results for two minutes of data for each of the scenarios. Selected frames from such videos are presented in Fig. 7, and the corresponding videos can be seen on the companion website. In the images and videos, the tracker output is displayed for each person as an ellipse of distinct tone. Inferred speaking activity is shown as a double ellipse with contrasting tones.

From Fig. 6(a) and (c), we can observe that the continuous estimation of 3-D location is quite stable in cases where speakers are seated, and the average error remains low (on average 12 cm for stationary, and 22 cm for overlap, as seen in Table II). These errors are partially due to the fact that the tracker estimate in each camera view corresponds to the center of a person's head, which introduces errors because, in strict terms, the two head centers do not correspond to the same physical 3-D point, and also because they do not correspond to the mouth center. The overlap case is clearly more challenging than the stationary one.

TABLE II
TRACKING RESULTS. 3-D ERROR BETWEEN GROUND TRUTH AND AUTOMATIC METHODS. THE STANDARD DEVIATION IS IN BRACKETS

	Error (m)			
Scenario	Audio-visual	Audio-only Method-average	Audio-only Method-minimum	Video-only
Stationary	0.12 (0.022)	0.23 (0.126)	0.20 (0.100)	0.10 (0.018)
Moving	0.38 (0.148)	0.84 (0.403)	0.70 (0.363)	–
Overlap	0.22 (0.014)	0.93 (0.932)	0.60 (0.468)	0.25 (0.017)

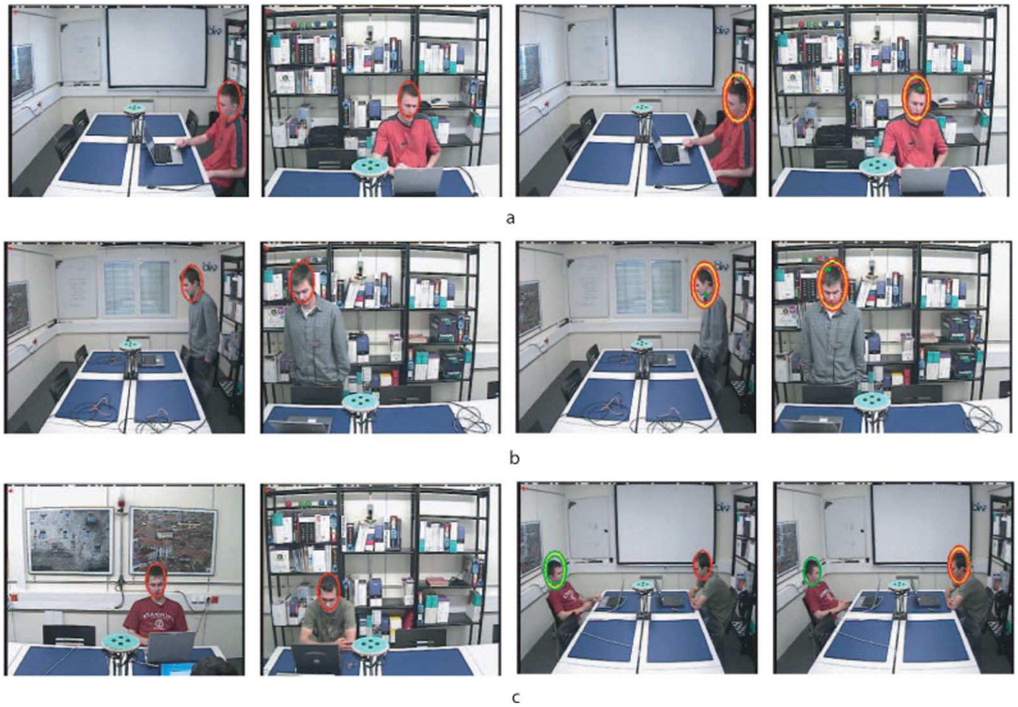

Fig. 7. (a) Tracking a single speaker in the stationary case, and (b) the moving case. (c) Tracking two speakers in the overlapping speech case. The speakers are tracked in each view and displayed with an ellipse. A "+" symbol indicates audio location estimate. A contrasting tone ellipse indicates when the speaker is active.

For the moving case, illustrated in Fig. 6(b), we observe an increased error (38 cm on average), which can be explained at least partially by the inaccuracy of the dynamical model, e.g., when the speaker stops and reverses the direction of motion, the tracker needs some time to adjust. This is evident in the corresponding video.

To validate our approach with respect to an audio-only algorithm, we also evaluated the results using directly the 3-D output of the speaker localization algorithm. Results are also shown in Table II, Figs. 6 and 7 and the website videos. In images and videos, the audio-only estimates are represented by "+" symbols.

Recall from Section V-B that the audio localization algorithm outputs between zero and three audio estimates per video frame. Using this information, we compute two types of errors. The first one uses the *average* Euclidean distance between the ground truth and the available audio estimates. The second one uses the *minimum* Euclidean distance between the ground truth and the automatic results, which explicitly considers the best (*a posteriori*) available estimate. While the first case can be seen as a fair, blind evaluation of the accuracy of location estimation, the second case can be seen as a best case scenario, in which a form of data association has been done for evaluation. As shown in Fig. 6, the audio-only estimates are discontinuous and are available only in approximately 60% of the frames. Errors are computed only on those frames for which there is at least one audio estimate. The results show that, in all cases, the performance obtained with audio-only information is consistently worse than that obtained with the multimodal approach, regarding both means and standard deviation. When the average Euclidean distance is used, performance degrades by almost 100% for the stationary case, and even more severely for the moving and overlap cases. Furthermore, while the best-case scenario results (minimum Euclidean distance) clearly reduce the errors for audio, due to the *a posteriori* data association, they nevertheless remain consistently worse than those obtained with the audio–visual approach. Importantly, compared to the audio–visual case, the reliability of the audio estimates (for both average and minimum) degrades more considerably when going from the single-speaker case to the concurrent-speakers case.

We also compared our algorithm with a variation of our multiperson tracker where only video observations were used (obviously in this case, the tracker cannot infer speaking activity). All other parameters of the model remained the same. In this case, the localization performance was similar to the audio–visual case for the stationary and overlapping speech cases, as indicated in Table II. However, the performance of the video-only tracker degraded in the case of moving speaker, as the tracker was affected by clutter (the bookshelf in the background) and lost track in some sequences (which is the reason why results for

TABLE III
SNRE RESULTS

Signal	SNRE (dB) Stationary	Moving	Overlap
Headset	24.8	23.4	17.4
Lapel	16.2	15.2	11.2
Audio Beamformer	13.1	11.8	5.2
Audio Beamformer + Post-filter	14.4	12.2	5.6
AV Beamformer	15.3	13.4	6.7
AV Beamformer + Post-filter	16.8	13.6	10.1

TABLE IV
ADAPTATION AND TEST DATA DESCRIPTION

Scenario	No. of sentences Adaptation	Testing
Stationary	60	100
Moving	30	48
Overlap	24	46

TABLE V
SPEECH RECOGNITION RESULTS

Signal	WER (%) Stationary	Moving	Overlap
Headset	21.3	19.3	42.9
Lapel	27.9	24.4	49.7
Distant Microphone	37.5	37.6	95.3
Audio Beamformer	31.3	33.3	81.7
Audio Beamformer + Post-filter	32.8	34.6	80.3
AV Beamformer	26.8	28.5	69.4
AV Beamformer + Post-filter	26.3	29.4	56.6

this case are not reported in Table II). Overall, compared to the audio-only and video-only approaches, the multimodal tracker yields clear benefits.

C. Speech Enhancement and Recognition Experiments

To assess the noise reduction and evaluate the effectiveness of the microphone array in acquiring a "clean" speech signal, the segmental signal-to-noise ratio (SNR) is calculated. To normalize for different levels of individual speakers, all results are quoted with respect to the input on a single table-top microphone, and hence represent the SNR enhancement (SNRE). These results are shown in Table III.

Speech recognition experiments were performed to evaluate the performance of the various scenarios. The number of sentences for adaptation and test data are shown in Table IV. Adaptation data was taken from the DEV set and test data was taken from the EVAL set. Adaptation data was matched to the corresponding testing channel condition. In MLLR adaptation, a static two-pass approach was used, where in the first pass, a global transformation was performed, and in the second pass, a set of specific transforms for speech and silence models were calculated. The MLLR transformed means are used as the priors for the MAP adaptation. All the results are scenario-specific, due to the different amounts of adaptation and test data. Table V shows the speech recognition results after adaptation.

In the following, we summarize the discussion regarding the speech enhancement and speech recognition experiments.

Headset, lapel, and distant microphones: As can be seen from Tables III and V, as expected for all the scenarios (stationary, moving, and overlap speech) and all the testing conditions (headset, lapel, distant, audio beamformer, audio beamformer + postfilter, audio–visual (AV) beamformer, AV beamformer + post-filter), the headset speech has the highest SNRE, which in turn results in the best speech recognition performance. Note that the obtained WER corresponds to the typical recognition results with the 20k WSJ task comparable with the 20.5% obtained with the baseline system described in the previous section. The headset case can thus be considered as the baseline for all the results from the other channels to be compared. The lapel microphone offers the next best performance, due to its close proximity (around 8 cm.) to the mouth of the speaker. Regarding the distant microphone signal, the WER obtained in this case is due to the greater susceptibility to room reverberation and low SNR, because of its distance (around 80 cm.) from the desired speaker. In all cases, the severe degradation in SNRE and WER for the overlap case compared to the single speaker case is self-evident, although obviously headset is the most robust case.

Audio-only: The audio beamformer and audio beamformer + postfilter perform better than the distant microphone for all scenarios, for both SNRE and WER. It can be observed that the postfilter helps in acquiring a better speech signal than the beamformer. However, the SNR and WER performances are in all cases inferior when compared to the headset and lapel microphone cases. This is likely due to the fact that the audio estimates are discontinuous and not available all the time, are affected by audio clutter due to laptops and computers in the meeting room, and are highly vulnerable to the room reverberation.

Audio–visual: From Tables III and V, it is clear that the AV beamformer and AV beamformer + postfilter cases perform consistently better than the distant microphone and audio-only systems for both SNRE and WER. In the single stationary speaker scenario, the AV beamformer + postfilter performs better than lapel, suggesting that the postfilter helps in speech enhancement without substantially distorting the beamformed speech signal. This is consistent with earlier studies which have shown that recognition results from beamformed channels can be comparable or sometimes better than lapel microphones [45]. In the overlapping speech scenario, the postfilter specially designed to handle overlapping speech is effective in reducing the crosstalk speech. The postfilter significantly improved the beamformer output, getting close to the lapel case in terms of SNRE, but less so in terms of WER. It can also be observed that there is no clear benefit to the postfilter over the beamformer in the moving single-speaker scenarios. Some examples of enhanced speech are available on the companion website.

D. Limitations and Future Work

Our system has a number of limitations. The first one refers to the audio–visual tracking system. As illustrated by the video-only results, the visual features can sometimes fail when a combination of background clutter and differences between the predicted dynamics and the real motion occur, which results in tracking loss. We are considering the inclusion of stronger cues about human modeling (e.g., face detectors), or features derived from background modeling techniques to handle these cases. However, their introduction needs to be handled with care, as one of the advantages of our approach is its ability to model variations of head pose and face appearance without needing a heavy model training phase with large number of samples (e.g., required for face detectors), or background adaptation methods. The second limitation comes from the use of a small microphone array, which might not be able to provide as accurate location estimates as a large array. However, small microphone arrays are beneficial in terms of deployment and processing, and the location accuracy is not affected so much in small spaces like the one used for our experiments. Further research could also investigate more sophisticated methods to update the beamformer filters based on the tracked location, or methods for achieving a closer integration between the speech enhancement and recognition stages.

IX. Conclusion

This paper has presented an integrated framework for speech recognition from data captured by an audio–visual sensor array. An audio–visual multiperson tracker is used to track the active speakers with high accuracy, which is then used as input to a superdirective beamformer. Based on the location estimates, the beamformer enhances the speech signal produced by a desired speaker, attenuating signals from the other competing sources. The beamformer is followed by a novel post-filter which helps in further speech enhancement by reducing the competing speech. The enhanced speech is finally input into a speech recognition module.

The system has been evaluated on real meeting room data for single stationary speaker, single moving speaker, and overlapping speakers scenarios, comparing in each case various single channel signals with the tracked, beamformed, and postfiltered outputs. The results show that, in terms of SNRE and WER, our system performs better than a single table-top microphone, and is comparable in some cases to lapel microphones. The results also show that our audio–visual-based system performs better than an audio-only system. This shows that accurate speaker tracking provided by a multimodal approach was beneficial to improve speech enhancement, which resulted in improved speech recognition performance.

Acknowledgment

The authors would like to thank G. Lathoud (IDIAP) for help with audio source localization experiments, D. Moore (CSIRO) for help with the initial speech enhancement experiments, J. Vepa (IDIAP) for his support with the speech recognition system, M. Lincoln (CSTR, University of Edinburgh) for the collaboration in designing the MC-WSJ-AV corpus, S. Ba (IDIAP) for his support with the audio–visual sensor array calibration, and B. Crettol (IDIAP) for his support to collect the data. They would also like to thank all the participants involved in the recording of the corpus.

References

[1] G. Abowd *et al.*, "Living laboratories: The future computing environments group at the Georgia Institute of Technology," in *Proc. Conf. Human Factors in Comput. Syst. (CHI)*, Hague, Apr. 2000, pp. 215–216.

[2] F. Asano *et al.*, "Detection and separation of speech event using audio and video information fusion," *J. Appl. Signal Process.*, vol. 11, pp. 1727–1738, 2004.

[3] R. Baeza-Yates and B. Ribeiro-Neto*, Modern Information Retrieval*. New York: ACM, 1999.

[4] M. Beal, H. Attias, and N. Jojic, "Audio-video sensor fusion with probabilistic graphical models," in *Proc. Eur. Conf. Comput. Vision (ECCV)*, Copenhagen, May 2002.

[5] J. Bitzer, K. S. Uwe, and K. Kammeyer, "Theoretical noise reduction limits of the generalized sidelobe canceller (GSC) for speech enhancement," in *Proc. Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 1999, vol. 5, pp. 2965–2968.

[6] R. Brunelli *et al.*, "A generative approach to audio–visual person tracking," in *Proc. CLEAR Evaluation Workshop*, Southampton, U.K., Apr. 2006, pp. 55–68.

[7] N. Checka, K. Wilson, M. Siracusa, and T. Darrell, "Multiple person and speaker activity tracking with a particle filter," in *Proc. Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Montreal, QC, Canada, May 2004, pp. V-881–V-884.

[8] R. Chellapa, C. Wilson, and A. Sirohey, "Human and machine recognition of faces: A survey," *Proc. IEEE*, vol. 83, no. 5, pp. 705–740, May 1995.

[9] Y. Chen and Y. Rui, "Real-time speaker tracking using particle filter sensor fusion," *Proc. IEEE*, vol. 92, no. 3, pp. 485–494, Mar. 2004.

[10] S. M. Chu, E. Marcheret, and G. Potamianos, "Automatic speech recognition and speech activity detection in the chil smart room," in *Proc. Joint Workshop Multimodal Interaction and Related Machine Learning Algorithms (MLMI)*, Edinburgh, U.K., Jul. 2005, pp. 332–343.

[11] R. K. Cook, R. V. Waterhouse, R. D. Berendt, S. Edelman, and M. C. Thompson, Jr, "Measurement of correlation coefficients in reverberant sound fields," *J. Acoust. Soc. Amer.*, vol. 27, pp. 1072–1077, 1955.

[12] H. Cox, R. Zeskind, and M. Owen, "Robust adaptive beamforming," *IEEE Trans. Acoust., Speech. Signal Process.*, vol. ASSP-35, no. 10, pp. 1365–1376, Oct. 1987.

[13] H. Cox, R. Zeskind, and I. Kooij, "Practical supergain," *IEEE Trans. Acoust., Speech. Signal Process.*, vol. ASSP-34, no. 3, pp. 393–397, Jun. 1986.

[14] J. Crowley and P. Berard, "Multi-modal tracking of faces for video communications," in *Proc. Conf. Comput. Vision Pattern Recognition (CVPR)*, San Juan, Puerto Rico, Jun. 1997, pp. 640–645.

[15] J. DiBiase, "A high-accuracy, low-latency technique for talker localization in reverberant environments," Ph.D. dissertation, Brown Univ., Providence, RI, 2000.

[16] J. DiBiase, H. Silverman, and M. Brandstein, "Robust localization in reverberant rooms," in *Microphone Arrays*. New York: Springer, 2001, vol. 8, pp. 157–180.

[17] G. W. Elko, "Superdirectional microphone arrays," in *Acoustic Signal Processing for Telecommunication*, S. Gay and J. Benesty, Eds. Norwell, MA: Kluwer, 2000, ch. 10, pp. 181–237.

[18] A. Doucet, N. de Freitas, and N. Gordon*, Sequential Monte Carlo Methods in Practice*. New York: Springer-Verlag, 2001.

[19] J. Fisher, T. Darrell, W. T. Freeman, and P. Viola, "Learning joint statistical models for audio–visual fusion and segregation," in *Proc. Neural Inf. Process. Syst. (NIPS)*, Denver, CO, Dec. 2000, pp. 772–778.

[20] D. Gatica-Perez, G. Lathoud, I. McCowan, and J.-M. Odobez, "A mixed-state i-Particle filter for multi-camera speaker tracking," in *Proc. IEEE Conf. Comput. Vision, Workshop on Multimedia Technologies for E-learning and Collaboration(ICCV-WOMTEC)*, Nice, France, Oct. 2003.

[21] D. Gatica-Perez, G. Lathoud, J.-M. Odobez, and I. McCowan, "Multimodal multispeaker probabilistic tracking in meetings," in *Proc. IEEE Conf. Multimedia Interfaces (ICMI)*, Trento, Italy, Oct. 2005, pp. 183–190.

[22] D. Gatica-Perez, G. Lathoud, J.-M. Odobez, and I. McCowan, "Audio-visual probabilistic tracking of multiple speakers in meetings," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 2, pp. 601–616, Feb. 2007.

[23] J.-L. Gauvain and C.-H. Lee, "Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains," *IEEE Trans. Acoust., Speech. Signal Process.*, vol. 2, no. 2, pp. 291–298, Apr. 1994.

[24] S. M. Griebel and M. S. Brandstein, "Microphone array source localization using realizable delay vectors," in *Proc. IEEE Workshop Applicat. Signal Process. Audio Acoust. (WASPAA)*, New York, Oct. 2001, pp. 71–74.

[25] T. Hain *et al.*, "The development of the AMI system for the transcription of speech in meetings," in *Proc. Joint Workshop Multimodal Interaction and Related Mach. Learn. Algorithms (MLMI)*, Edinburgh, U.K., Jul. 2005, pp. 344–356.

[26] R. Hartley and A. Zisserman*, Multiple View Geometry in Computer Vision*, 2nd ed. Cambridge, U.K.: Cambridge Univ. Press, 2001.

[27] M. Isard and A. Blake, "CONDENSATION: Conditional density propagation for visual tracking," *Proc. Int. J. Comput. Vision*, vol. 29, no. 1, pp. 5–28, 1998.

[28] B. Kapralos, M. Jenkin, and E. Milios, "Audio-visual localization of multiple speakers in a video teleconferencing setting," *Int. J. Imaging Syst. Technol.*, vol. 13, pp. 95–105, 2003.

[29] N. Katsarakis *et al.*, "3D audiovisual person tracking using Kalman filtering and information theory," in *Proc. CLEAR Evaluation Workshop*, Southampton, U.K., Apr. 2006, pp. 45–54.

[30] Z. Khan, T. Balch, and F. Dellaert, "An MCMC-based particle filter for tracking multiple interacting targets," in *Proc. Eur. Conf. Comput. Vision (ECCV)*, Prague, May 2004, pp. 279–290.

[31] J. Kleban and Y. Gong, "HMM adaptation and microphone array processing for distant speech recognition," in *Proc. Int. Conf. Acoust. , Speech, Signal Process. (ICASSP)*, Istanbul, Turkey, Jun. 2000, pp. 1411–1414.

[32] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *IEEE Trans. Acoust., Speech. Signal Process.*, vol. ASSP-24, no. 4, pp. 320–327, Aug. 1976.

[33] H. Krim and M. Viberg, "Two decades of array signal processing research: The parametric approach," *IEEE Signal Process. Mag.*, vol. 13, no. 4, pp. 67–94, Jul. 1996.

[34] G. Lathoud and I. McCowan, "A sector-based approach for localization of multiple speakers with microphone arrays," in *Proc. ISCA Workshop Statistical and Perceptual Audio Process. (SAPA)*, Jeju, Korea, Oct. 2004.

[35] C. J. Leggetter and P. C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," *Comput. Speech Lang.*, vol. 9, no. 2, pp. 171–185, 1995.

[36] J. S. Liu*, Monte Carlo Strategies in Scientific Computing*. New York: Springer-Verlag, 2001.

[37] M. Lincoln, I. McCowan, J. Vepa, and H. K. Maganti, "The multi-channel Wall Street Journal audio–visual corpus (MC-WSJ-AV): Specification and initial experiments," in *IEEE Autom. Speech Recognition Understanding Workshop (ASRU)*, San Juan, Puerto Rico, Dec. 2005, pp. 357–362.

[38] K. S. Uwe, J. Bitzer, and C. Marro, "Post-filtering techniques," in *Microphone Arrays*. New York: Springer, 2001, vol. 3, pp. 36–60.

[39] M. Wolfel, K. Nickel, and J. McDonough, "Microphone array driven speech recognition: Influence of localization on the word error rate," in *Proc. Joint Workshop Multimodal Interaction and Related Mach. Learn. Algorithms (MLMI)*, Edinburgh, U.K., Jul. 2005, pp. 320–331.

[40] C. Marro, Y. Mahieux, and K. U. Simmer, "Analysis of noise reduction and dereverberation techniques based on microphone arrays with postfiltering," *IEEE Trans. Speech Audio Process.*, vol. 6, no. 3, pp. 240–259, May 1998.

[41] I. McCowan and H. Bourlard, "Microphone array post-filter based on noise field coherence," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 6, pp. 709–716, Nov. 2003.

[42] I. McCowan, M. Hari-Krishna, D. Gatica-Perez, D. Moore, and S. Ba, "Speech acquisition in meetings with an audio–visual sensor array," in *Proc. IEEE Int. Conf. Multimedia (ICME)*, Amsterdam, The Netherlands, Jul. 2005, pp. 1382–1385.

[43] J. Meyer and K. U. Simmer, "Multi-channel speech enhancment in a car environment using Wiener filtering and spectral subtraction," in *Proc. Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Munich, Germany, Apr. 1997, pp. 1167–1170.

[44] N. Morgan, D. Baron, J. Edwards, D. Ellis, D. Gelbart, A. Janin, T. Pfau, E. Shriberg, and A. Stolcke, "The meeting project at ICSI," in *Proc. Human Lang. Technol. Conf.*, San Diego, CA, Mar. 2001, pp. 1–7.

[45] D. Moore and I. McCowan, "Microphone array speech recognition: Experiments on overlapping speech in meetings," in *Proc. Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Hong Kong, Apr. 2003, pp. V-497–V-500.

[46] K. Nickel, T. Gehrig, H. K. Ekenel, J. McDonough, and R. Stiefelhagen, "An audio–visual particle filter for speaker tracking on the CLEAR'06 evaluation dataset," in *Proc. CLEAR Evaluation Workshop*, Southampton, U.K., Apr. 2006, pp. 69–80.

[47] J. G. Fiscus, N. Radde, J. S. Garofolo, A. Le, J. Ajot, and C. Laprun, "The rich transcription 2005 spring meeting recognition evaluation," in *Proc. NIST MLMI Meeting Recognition Workshop*, Edinburgh, U.K., Jul. 2005, pp. 369–389.

[48] M. Omologo, M. Matassoni, P. Svaizer, and D. Giuliani, "Microphone array based speech recognition with different talker-array positions," in *Proc. Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Munich, Germany, Apr. 1997, pp. 227–230.

[49] T. R. al, "WSJCAM0: A British English speech corpus for large vocabulary continuous speech recognition," in *Proc. Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Detroit, MI, Apr. 1995, pp. 81–84.

[50] S. Roweis, "Factorial models and refiltering for speech separation and denoising," in *Proc. Eurospeech Conf. Speech Commun. Technol. (Eurospeech-2003)*, Geneva, Switzerland, Sep. 2003, pp. 1009–1012.

[51] D. Sturim, M. Brandstein, and H. Silverman, "Tracking multiple talkers using microphone-array measurements," in *Proc. Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Munich, Germany, Apr. 1997, pp. 371–374.

[52] B. D. V. Veen and K. M. Buckley, "Beamforming: A versatile approach to spatial filtering," *IEEE Acoust., Speech, Signal Process. Mag.*, vol. 5, no. 2, pp. 4–24, Apr. 1988.

[53] J. Vermaak, M. Gagnet, A. Blake, and P. Perez, "Sequential Monte Carlo fusion of sound and vision for speaker tracking," in *Proc. Int. Conf. Comput. Vision (ICCV)*, Vancouver, BC, Canada, Jul. 2001, pp. 741–746.

[54] A. Waibel, T. Schultz, M. Bett, R. Malkin, I. Rogina, R. Stiefelhagen, and J. Yang, "Smart: The smart meeting room task at ISL," in *Proc. Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Hong Kong, Apr. 2003, pp. IV-752–IV-754.

[55] D. Ward and R. Williamson, "Particle filter beamforming for acoustic source localization in a reverberant environment," in *Proc. Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Orlando, FL, May 2002, pp. 1777–1780.

[56] S. J. Young *et al.*, "Multilingual large vocabulary speech recognition: The European SQUALE project," *Comput. Speech Lang.*, vol. 11, no. 1, pp. 73–89, 1997.

[57] R. Zelinski, "A microphone array with adaptive post-filtering for noise reduction in reverberant rooms," in *Proc. Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, New York, Apr. 1988, pp. 2578–2581.

[58] Z. Zhang, "Flexible camera calibration by viewing a plane from unknown orientations," in *Proc. Int. Conf. Computer Vision (ICCV)*, Kerkyra, Greece, Sep. 1999, pp. 666–673.

[59] D. Zotkin, R. Duraiswami, and L. Davis, "Multimodal 3-D tracking and event detection via the particle filter," in *Proc. Int. Conf. Comput. Vision, Workshop on Detection and Recognition of Events in Video (ICCV-EVENT)*, Vancouver, BC, Canada, Jul. 2001, pp. 20–27.

Hari Krishna Maganti (S'05) graduated from the Institute of Electronics and Telecommunication Engineers, New Delhi, India, in 1997, received the M.E. degree in computer science and engineering from University of Madras, Madras, India, in 2001, and the Ph.D. degree in Engineering Science and Computer Sciences from University of Ulm, Ulm, Germany, in 2007.

His Ph.D. work included two years of research in multimedia signal processing at IDIAP Research Institute, Martigny, Switzerland. Apart from academic research, he has been involved in industry for more than three years working across different application domains. His primary research interests include audio–visual tracking and speech processing, particularly speech enhancement and recognition, speech/nonspeech detection, and emotion recognition from speech.

Daniel Gatica-Perez (S'01–M'02) received the B.S. degree in electronic engineering from the University of Puebla, Puebla, Mexico, in 1993, the M.S. degree in electrical engineering from the National University of Mexico, Mexico City, in 1996, and the Ph.D. degree in electrical engineering from the University of Washington, Seattle, in 2001.

He joined the IDIAP Research Institute, Martigny, Switzerland, in January 2002, where he is now a Senior Researcher. His interests include multimedia signal processing and information retrieval, computer vision, and statistical machine learning.

Dr. Gatica-Perez is an Associate Editor of the IEEE TRANSACTIONS ON MULTIMEDIA.

Iain McCowan (M'97) received the B.E. and B.InfoTech. degrees from the Queensland University of Technology (QUT), Brisbane, Australia, in 1996 and the Ph.D. degree with the research concentration in speech, audio and video technology at QUT in 2001, including a period of research at France Telecom, Lannion.

He joined the IDIAP Research Institute, Martigny, Switzerland, in April 2001, as a Research Scientist, progressing to the post of Senior Researcher in 2003. While at IDIAP, he worked on a number of applied research projects in the areas of automatic speech recognition and multimedia content analysis, in collaboration with a variety of academic and industrial partner sites. From January 2004, he was Scientific Coordinator of the EU AMI (Augmented Multi-Party Interaction) project, jointly managed by IDIAP and the University of Edinburgh. He joined the CSIRO eHealth Research Centre, Brisbane, in May 2005 as Project Leader in multimedia content analysis and is a part-time Research Fellow with the QUT Speech and Audio Research Laboratory.