Augmented **M**ulti-party **I**nteraction
http://www.amiproject.org

Augmented **M**ulti-party **I**nteraction with **D**istance **A**ccess
http://www.amidaproject.org

State-of-the-art overview

Localization and Tracking of Multiple Interlocutors with Multiple Sensors

Updated version 23.01.2007

1 Introduction

The automatic analysis of meetings recorded in multi-sensor rooms is an emerging research field in various domains, including audio and speech processing, computer vision, human-computer interaction, and information retrieval [45, 86, 64, 81, 21, 61, 92]. Analyzing meetings poses a diversity of technical challenges, and opens doors to a number of relevant applications, including automatic structuring and indexing of meeting collections, and facilitation of remote meetings.

In the context of meetings, detecting, localizing, and tracking people and their speaking activity are crucial for enriching the interactive experience between participants of a meeting, on a single site as well as on multiple sites, e.g. videoconferencing. The localization and tracking tasks play fundamental roles in two areas. The first one is media processing: speaker location is useful to select or steer a camera as part of a visualization or production model, to enhance the audio stream via microphone-array beamforming for speech recognition, to provide accumulated information for person identification, and to recognize location-based events (e.g. a presentation). The second one is human interaction analysis: social psychology has highlighted the role of non-verbal behavior (e.g. gaze and facial expressions) in interactions, and the correlation between speaker turn patterns and aspects of the behavior of a group [63]. Extracting cues to identify such multimodal behaviors requires reliable speaker localization and tracking.

Although the above tasks are facilitated in meetings by the constraints of the physical space and the expected type of human activities, they still pose several challenges. The interactive nature of meetings involves spontaneous multi-party speech, which contains highly dynamic patterns of speaker turns, including short speech utterances, non-linear human motion, partial and total visual occlusion, and multiple sources (multiple overlapping speakers and/or background noise sources). The meeting room environment is thus highly dynamical, and in order to develop useful applications for enhanced user experience *during* meetings (e.g. online and realtime) and *after* meetings (e.g. automatic summarization or queries from a user), robust and computationally tractable methods that can cope with the multisource aspect as well as the highly dynamical aspect are necessary. Ideally, such methods should address, in principled ways, the need for fusion of perceptual data (e.g. multiple audio and visual sources), in order to exploit the modalities' redundancy and complementarity, and the need for accurate descriptions of the interactive, multiperson processes that meetings contain (e.g. representing the dynamic status of each individual, while accounting for the constraints introduced by their interaction).

This report presents an overview of existing work on localization and tracking of multiple interlocutors with multiple sensors. Rather than being exhaustive, the report attempts to provide a non-expert reader with pointers to what the authors regard as representative work in the domain, and to present a succint discussion of the advantages and limitations of such methods, including the ones that have been developed as part of the AMI project. The *multisource* context will be the focus throughout. Additionally, given that the emphasis of the review is on localization and tracking of talking people, we limit the review to audio-only and audio-visual (AV) methods.

The report is organized as follows. Section 2 reviews existing work on localization and tracking with audio sensors. Section 3 does so for work using audio-visual sensors. Finally, Section 4 discusses available resources (i.e., data and related annotations) in this domain.

2 Localization and tracking with audio sensors

The speed of sound in the air being finite and relatively low in an indoor environment (around 342 m/s), most practical audio localization/tracking applications rely on time asynchrony between the waves arriving at multiple microphones in multiple locations, called microphone arrays. A three-fold inverse problem then arises: that of inferring, from the slight differences between the recorded signals, the number of active speakers at any given time (detection), their instantaneous positions (localization) and their spatio-temporal trajectories over time (tracking).

These "slight differences" are tightly linked to the geometrical placement of the microphones. More precisely, the differences are usually measured from three non-exclusive viewpoints:

1. Time asynchrony: the time of flight of the acoustic waves from mouth to microphone is different for different microphones, due to their different placements. Audio localization/tracking methods relying on time asynchrony usually require precise knowledge of the microphone array's geometry. However, they do not require any particular knowledge about the room, and are thus the most developed in terms of practical applications. Omnidirectional microphones are used in most cases. Typical geometries include Uniform Linear Arrays (ULA) and Uniform Circular Arrays (UCA): a finite number of microphones equally spaced along a line or a circle, respectively. However, a solution particularly designed for meeting rooms is the Huge Microphone Array (HMA) including a large number of microphones on a wall [76].

2. Impulse response: for a given mouth location, the path travelled by the sound to the various microphones will vary. Hence, the impulse response will vary as well. Assuming the impulse response characteristics of the room to be perfectly known beforehand (calibration), it is possible to deduce the position of the speaker. However, the tedious calibration step is often undesirable in practical application (e.g. portable videoconferencing systems), so the impulse responses need to be estimated in an online, automatic fashion. This task is also called blind Multiple Inputs Multiple Outputs (MIMO) channel identification [16, 11], where geometrical knowledge of the microphone array is not required. The task is tightly linked to Blind Source Separation approaches [12]. Solving this problem amounts to retrieve a complete model of the meeting room, which would permit not only to locate but also to separate the various signals at the same time. This problem is still difficult and open to research. A preliminary test of such a method can be found in [13].

3. Microphone channel: recently, it was proposed to use several directional microphones placed at the same location, but oriented towards different directions [60]. The direction-dependent transfer function of each microphone is assumed to be known, so that the speaker location can be reconstructed. This solution can also be combined with the first group of approaches [72].

In the following we focus on the first group of solutions, for which [9, 44] provide comprehensive introductions. These methods are typically linked, directly or indirectly, to the following observation: if two signals $x_1(t)$ and $x_2(t) = x_1(t - \tau_{12})$ are received by two microphones, with a relative delay τ_{12}, the cross-correlation function $f(\tau) = \int_t x_1(t)x_2(t-\tau)dt$ will have a maximum at $\tau = \tau_{12}$. We first briefly mention the detection issue within the context of source localization, then examine the various instantaneous source localization methods. Finally, tracking of speaker trajectories over time is presented.

2.1 Detection

Since the speech of a given person is sporadic, it is needed *not* to estimate any speaker location during silence. Traditional Voice Activity Detectors (VADs) typically use single channel features, such as energy and zero-crossing rate. Although well adapted to single channel tasks such as automatic speech recognition, they are suboptimal in terms of localization precision [49]. Thus, they are not necessarily adapted to the task of acoustic source localization. Alternative methods that rely on the cross-correlation between channels can be found in [17, 51, 52].

2.2 Localization

The usual meaning of "localization" is to identify the location of the various active speech sources in physical space, from a short time frame on which speech is considered as stationary (typically 20 to 30 ms). As mentioned above, methods based on the asynchrony between the various microphones rely on the cross-correlation between those signals. They can be divided into two types: time delay of arrival (TDOA), and direct methods.

The TDOA methods consist in first estimating the time delays between each pair of microphones, and then deriving the location of the sources from geometrical considerations, e.g. using the Linear Intersection algorithm [8]. The main bottleneck is the time delay estimation (TDE) step, which may be affected by reverberations. A practical improvement can be obtained by modifying the cross-correlation function, e.g. using the generalized cross correlation phase transform (GCC-PHAT) [43]. In the case of multiple sources and one microphone pair, an alternative method for the estimation of multiple time-delays is the Adaptive Eigenvalue Decomposition Algorithm [6]. However, in the general case of multiple microphone pairs, multiple sources, and multiple sound paths (reverberations), it is not obvious how to pair the various time-delays observed at the various pairs of microphones in order to deduce the exact location of the acoustic sources.

The direct methods avoid this bottleneck by directly inferring the source(s) locations from the measured signal. They can be divided into two groups: Coherent Signal Subspace Processing (CSSP), and beamforming. CSSP methods [87, 19] are extensions of narrowband methods originated in the fields of radar and communications [73]. Although they allow in theory to estimate jointly the number of sources and their locations, they suffer from sensitivity to reverberant environments and/or need sufficient amounts of data.

Beamforming localization methods, also known as Steered Response Power (SRP) methods, are a reasonable alternative that does not rely on strong room knowledge/modelling assumptions. The idea is to estimate the power at any location in space by compensating for the corresponding delays between the signals ("steering" the array) [44]. Multiple simultaneous sources will be reflected by multiple maxima across the search space. However, reverberations will also appear as maxima ("virtual sources"). A partial solution to this issue is to combine the flexibility of SRP methods with the robustness to reverberations of PHAT: this is known as SRP-PHAT [22]. In general, the main drawback of SRP methods is that the search space can be large (e.g. the whole room in the case of meetings). Recently, an approach was proposed that discretizes the search space into volumes, and reduces the search to "active" volumes only [25, 26]. In such a framework, localization amounts to determine whether there is an acoustic source present within each of a predetermined, finite number of volumes. However, spatial resolution and interference between the signals of the different speakers may be an issue. Indeed, using the whole spectrum to locate multiple sources leads to unnecessary noise in the location estimation. In other words, spectral data from all sources is used to locate a given source, thus inducing an unnecessary bias in the location estimate. A possible way to address these issues is the "sparsity assumption" in the frequency domain [51, 52]. This assumption is derived from

statistical observations on human speech [71], and simply means that within a frequency bin, only one speech source is dominant in terms of magnitude, while all other sources can be neglected.

In practice, although CSSP methods and, more recently, MIMO/BSS methods have seen very promising developments, they are still not as practically effective as the SRP methods. One drawback of SRP methods is that number of active sources and reverberations are not determined automatically. However, this issue can be efficiently addressed by clustering/tracking methods.

2.3 Tracking

Tracking can be viewed as the task of filtering instantaneous location estimates provided by the methods mentioned above. The Kalman filter [39, 90] assumes dynamics to be linear and Gaussian. These assumptions become an issue when dealing with human motion (non linearities such as sharp turns). Furthermore, in spontaneous speech, utterances are short (typically less than a second), speaker changes often, and overlaps represent a non-negligible portion of speech [75].

The Extended Kalman Filter (EKF) was proposed to accomodate non-linear dynamics through a linearization step [79], however it is known to be practically difficult to tune its parameters [37]. More recently the Unscented Kalman Filter (UKF) was proposed to avoid this linearization step and accomodate non-Gaussian measurement noise sources [38, 37, 53]. For a recent application of the UKF to acoustic source localization, see [24]. However, these approaches may encounter difficulties when dealing with spontaneous speech, which is both highly changing in space (speaker changes) and sporadic over time (short utterances).

As an alternative, Sequential Monte-Carlo (SMC) methods, also known as Particle Filtering (PF), approximate the optimal Bayesian filter by representing probability distributions through a finite set of particles [33, 23]. For a state-space model, a PF recursively approximates the filtering distribution of states given observations using a dynamical model, an observation model, and sampling techniques, by predicting candidate configurations and measuring their likelihood, in a process that amounts to random search in the configuration space. Applications to single acoustic source localization and tracking can be found in [83, 88, 89], and a comprehensive review in [54]. However, the fast-changing speaker turns encountered in spontaneous multi-party speech requires either specific multisource models [46] or adapting the single-source model to "switching between speakers" situations [55]. Estimating the number of active speech sources is still an issue, tightly linked to the data association issue. Although Particle Filters can model multiple objects via multi-modal distributions, deciding which modes are significant and which objects they belong to is an open issue. Moreover, when the number of active objects varies very often along time, complex birth/death rules are needed.

Recently, alternative approaches were proposed where the number of active speech sources need not be known [47]. An unsupervised, online approach called "short-term clustering" was proposed, that automatically groups location estimates that are close to each other in space and time, and separate those that are not. [48] demonstrates on real data that it can be directly applied to the task of spontaneous speech segmentation, without requiring any constraint from the participants. The resulting speech segmentation is very precise, even when multiple participants talk at the same time. In addition, "short-term clustering" is shown to allow for detection and solving of trajectory crossing issues.

3 Localization and tracking with audio-visual sensors

Localizing and tracking speakers in enclosed spaces using AV information has increasingly attracted attention in signal processing and computer vision [69, 34, 20, 67, 27, 84, 95, 2, 5, 18, 15], given

the complementary characteristics of each modality. Broadly speaking, the differences among existing works arise from the overall goal (tracking single vs. multiple speakers), the specific detection/tracking framework, and the AV sensor configuration. Much work has concentrated on the single-speaker case, assuming either single-person scenes [20, 67, 2], or multiperson scenes where only the location of the current speaker needs to be tracked [69, 34, 27, 84, 95, 5]. Many of these works used simple sensor configurations, i.e., one camera and a microphone pair [20, 67, 84, 5]. Other works have addressed the data fusion problem using multiple microphones and a single camera [28], and others have used multiple cameras, either non-calibrated [29] or fully calibrated [95, 65], given the fact that, while single cameras are useful for remote conferencing applications, multiperson conversational settings like meetings often call for the use of multiple sensors to cover the entire workspace (table, whiteboards, etc.). Among the existing techniques, probabilistic generative models based on exact [67] or approximate inference methods, both variational [5] and sampling-based [84, 95, 28, 29, 65], appear to be the most promising, given their principled formulation and demonstrated performance.

None of the above works, however, can handle the problem of continuously inferring, from audio and video data, the location and speaking status for multiple people in a realistic conversational setting. In fact, although audio-based multispeaker tracking and vision-based multiobject tracking have been studied for a few years as separate problems in signal processing [82, 70, 85, 51] and computer vision [36, 68, 93, 94], respectively, the AV multispeaker tracking problem has been studied only relatively recently, making use of more complex sensor configurations [21, 40, 77, 14, 15, 18, 4, 30]. Each of these works is briefly discussed in the following. For presentation purposes, we categorize the existing work as being either system-oriented or model-oriented, where the emphasis in the first case is on module integration, while in the second case is on a unifying mathematical formulation.

Regarding the first category, the work in [21] described a system based on a device that integrates a small circular microphone array and several calibrated cameras, whose views are merged into a panorama, The system, in which each person is tracked independently, consists of three modules: AV auto-initialization, using either a standard acoustic source localization algorithm or visual cues, visual tracking using a Hidden Markov Model (HMM), and tracking verification. The work in [40] described a non-probabilistic multispeaker detection algorithm using an omnidirectional camera (which has limitations of resolution) and a microphone array, calibrated with respect to each other. At each video frame, the method extracts skin-color blobs by traditional techniques, and then detects a sound source using standard beamforming on the small set of directions indicated by the skin-blob locations. The work in [77] described an AV multispeaker system, based on a stereo camera and a linear microphone array, consisting of three separate modules: stereo-based visual tracking of 3-D head location and pose for each person independently, estimation of the direction of arrival of the audio signal with the microphone array, and estimation of audio-visual synchronous activity. Two hypothesis tests are used to make independent decisions about the speaking activity and visual focus of the speakers, based on simple statistical models defined on the observations derived from each module. The work in [14] uses a number of standard techniques in separate modules that are later integrated into a system that estimates the location and identity of the meeting participants, and detects the current speaker, using a setup including four calibrated cameras (an omnidirectional camera located on the center of the meeting table, and four cameras located in the corners of the meeting room), and a 16-microphone array, located on one end of the table.

For the second category, a number of probabilistic generative models have been recently proposed for the task of simultaneously inferring location and speaking activity of multiple interlocutors. All of them are based on PF techniques [15, 18, 4, 30, 31], but differ in the choices of state space, dynamical model, observation model, and sampling technique. The work in [15] used two calibrated cameras and four linear sub-microphone arrays on a wall, and was based on the model first proposed

in [36], which defines a multi-person state-space where the number of people can vary over time. A full-body multi-person observation model was defined by two terms: one for video, derived from a pixelwise background substraction model, and one for audio, derived from a set of short-time Fourier transforms computed on each microphone's signal. The PF relied on basic importance sampling (IS), and so is likely to become rapidly inefficient as the number of people increases. The work in [18] used the same calibrated sensor setup as [21], and tracked multiple speakers with a set of independent PFs, one for each person. For sampling, each PF uses a mixture proposal distribution, in which the mixture components are derived from the output of single-cue trackers (based on audio, color, or shape information). This proposal distribution increases robustness in case of tracking failure in single modalities. Furthermore, each single-object observation model is assumed to be factorized over the various cues. The work in [4] uses a setup composed of a stereo camera and a circular 8-microphone array, and uses a basic PF to perform inference over a multi-person state space, assuming that the multi-object observation model can be factorized over participants. However, the approach was only applied to two-people scenes, likely due to the known limitations of the basic PF algorithm. The work in [30, 31], developed in the context of the AMI project, presents an approach in a meeting room consisting of three uncalibrated cameras covering the physical space with mostly non-overlapping fields-of-view, and a circular 8-microphone array placed on the center of the meeting table. The model uses a mixed-state, multi-object state-space, which integrates a pairwise person occlusion model through the addition of a Markov Random Field prior in the multi-object dynamic model. To address the problems of traditional PFs in handling the high-dimensional state space defined by the joint multi-person configurations, inference in this model is performed with a Markov Chain Monte Carlo particle filter (MCMC-PF), which results in high sampling efficiency [56, 42]. The model integrates audio-visual data through an observation model where audio observations are derived from a source localization algorithm, and visual observations are based on models of the shape and spatial structure of human heads. Overall, the model in [30, 31] has two advantages over [15, 18]. First, it explicitly incorporates a pairwise person interaction prior term, which is especially useful to handle person occlusion. Second, it uses an MCMC sampling technique, which allows to track several objects in a tractable manner (effectively close to the case of independent PFs), while preserving the rigorous joint state-space formulation.

An important initiative related to evaluation of audio-visual technologies for localization and tracking is the recent Workshop on Classification of Events, Actions and Relations (CLEAR), where audio-visual approaches were evaluated in the context of seminar and conference rooms to track single presenters on common data and using a common evaluation protocol [1, 10, 41, 66, 7, 32]. As representative examples, in [41], a 3D tracking with stand-alone video and audio trackers was combined using a Kalman filter. The work in [66] proposed an algorithm based on a particle filter approach to integrate acoustic source localization, person detection, and foreground segmentations using multiple cameras and multiple pairs of microphones. It was demonstrated that the specific audio-visual formulation yields greater tracking accuracy than a filter based on individual modalities. The reader is referred to the CLEAR workshop proceedings for details about all the approaches.

Some of the recent AMI work focused on integrating, improving and evaluating a system for hands-free speech recognition in meetings [62, 59, 58] based on an audio-visual sensor array, including the multi-modal approach for multi-person tracking [30, 31], and speech enhancement and recognition modules. As mentioned before, 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. However, with a few exceptions, speaker tracking research has been largely decoupled from microphone array speech recognition research. The work in [3] 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. In the work in [91], a particle filter that fuses audio from multiple large microphone arrays and video from multiple calibrated cameras was used in the context of seminar rooms, in which there is essentially one main speaker (the lecturer).

In the system in [59], audio is captured using a circular, table-top array of 8 microphones, and visual information is captured from 3 different camera views. Both audio and visual information are used to track the location of all active speakers in the meeting room [30, 31]. Speech enhancement is then achieved using microphone array beamforming followed by a novel post-filtering stage. The enhanced speech is finally input into a standard HMM recognizer system to 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. 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 multi-modal 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 outperforms single table-top microphones and is comparable to lapel microphone for all the scenarios, as measured by both 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 proved beneficial to both speech enhancement and recognition. An analysis of the effects of location accuracy on the recognition of overlapping speech is presented in [58].

4 Available Data Resources

Most of the research summarized in the previous two sections has been conducted over a number of non-standardized audio or audio-visual data sets, which vary from each other with respect to the specific sensor setup, the type of recorded situations, the structure of the data set, the type of existing annotations, and their degree of availability to others for research purposes. The community in this domain, however, has already acknowledged the considerable effort involved in collecting such data, and the need to rely in common evaluation procedures.

In the context of the AMI project, an audio-visual corpus, called AV16.3, was recorded and annotated, and reported in [50]. This corpus includes a high variety of scenarios, ranging from static, constrained cases, to dynamic and natural ones, with multiple seated or moving speakers in a meeting room. The sensors include two eight-microphone circular arrays on a table, and three cameras around the room. The calibration of the cameras allowed to reconstruct the ground-truth location of the mouth of each person with a 3-D error inferior to 1.2 cm. Overall, this data set should be interesting for both the audio and the vision communities, and is publicly available. A second corpus, called AV16.7, was recorded to evaluate the multi-person tracking task [78], and contains sequences including up to four people conversing and moving in the meeting room. Finally, an audio-visual corpus for speech recognition, called the Multi-Channel Wall Street Journal Audio-Visual (MC-WSJ-AV) corpus, was also recorded. The specification and structure of the full corpus are detailed in [57]. The corpus includes cases of single stationary speakers, single moving speakers, and stationary overlapping speakers. In the first scenario, the speaker reads out sentences from different positions within the meeting room. In the second one, the speaker moves between different positions while reading the sentences. Finally, in the third scenario, two speakers simultaneously read sentences from different positions within the room. Much of the data comprises non-native English speakers with different speaking styles and accents. The corpus is therefore suitable for research on both tracking and speech recognition.

Another important data resource is the one coordinated by NIST and the CHIL (Computers in the

Human Interaction Loop) european project through the CLEAR initiative, where data collected and annotated in the CHIL meeting and lecture rooms become available for purposes of common evaluation [80].

References

[1] A. Abad, C. Canton-Ferrer, C. Segura, J. L. Landabaso, D. Macho, J.R. Casas, J. Hernando, M. Pardas, C. Nadeu, "UPC Audio, Video and Multimodal Person Tracking Systems in the CLEAR Evaluation Campaign," in *Proc. Workshop on Classification of Events, Actions and Relations (CLEAR)*, Southampton, Apr. 2006.

[2] P. Aarabi and S. Zaky, "Robust sound localization using multi-source audiovisual information fusion," *Information Fusion*, vol. 3, no. 2, pp. 209–223, Sep. 2001.

[3] F. Asano et. al., "Detection and Separation of Speech Event using Audio and Video Information Fusion," *Journal of Applied Signal Processing*, Vol. 11, pp. 1727-1738, 2004.

[4] H. Asoh, F. Asano, T. Yoshimura, Y. Motomura, N. Ichimura, I. Hara, J. Ogata, and K. Yamamoto, "An application of a particle filter to bayesian multiple sound source tracking with audio and video information fusion," in *Proc. Int. Conf. on Information Fusion (IF)*, Stockholm,, Jun 2004.

[5] M. Beal, H. Attias, and N. Jojic, "Audio-video sensor fusion with probabilistic graphical models," in *Proc. European Conf. on Computer Vision (ECCV)*, May 2002.

[6] J. Benesty, "Adaptative eigenvalue decomposition algorithm for passive acoustic source localization," *Journal of the Acoustic Society of America*, vol. 107, no. 1, pp. 384–391, January 2000.

[7] K. Bernardin, T. Gehrig, R. Stiefelhagen, "Multi- and Single View Multiperson Tracking for Smart Room Environments," in *Proc. Workshop on Classification of Events, Actions and Relations (CLEAR)*, Southampton, Apr. 2006.

[8] M. Brandstein, *A Framework for Speech Source Localization Using Sensor Arrays*, Ph.D. thesis, Brown University, 1995.

[9] M. Brandstein and D. Ward, Eds., *Microphone Arrays*, Springer, 2001.

[10] R. Brunelli, A. Brutti, P. Chippendale, O. Lanz, M. Omologo, P. Svaizer, F. Tobia, "A Generative Approach to Audio-Visual Person Tracking," in *Proc. Workshop on Classification of Events, Actions and Relations (CLEAR)*, Southampton, Apr. 2006.

[11] H. Buchner, R. Aichner, and W. Kellerman, "Trinicon: A versatile framework for multichannel blind signal processing," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2004.

[12] H. Buchner, R. Aichner, and W. Kellermann, "Relation between blind system identification and convolutive blind source separation," in *Proc. HSCMA Workshop*, Piscataway, NJ, USA, Mar. 2005.

[13] H. Buchner, R. Aichner, J. Stenglein, H. Teutsch, and W. Kellermann, "Simultaneous localization of multiple sound sources using blind adaptive mimo filtering," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Philadelphia, PA, USA, Mar. 2005.

[14] C. Busso, S. Hernanz, C.-W. Chu, S.-I. Kwon, S. Lee, P. Georgiou, I. Cohen, and S. Narayanan, "Smart room: Participant and speaker localization and identification," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2005.

[15] N. Checka, K. Wilson, M. Siracusa, and T. Darrell, "Multiple person and speaker activity tracking with a particle filter," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Montreal, May 2004.

[16] J. Chen, J. Benesty, and A. Huang, "MIMO acoustic signal processing," Invited Talk, HSCMA Workshop, Mar. 2005.

[17] J.F. Chen and W. Ser, "Speech detection using microphone array," *Electronic Letters*, vol. 36, no. 2, Jan. 2000.

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

[19] E. Di Claudio and R. Parisi, "Multi-source localization strategies," in *Microphone Arrays*, M. Brandstein and D. Ward, Eds., chapter 9, pp. 181–201. Springer, 2001.

[20] R. Cutler and L. Davis, "Look who's talking: Speaker detection using video and audio correlation," in *Proc. IEEE Int. Conf. on Multimedia (ICME)*, New York, Jul. 2000.

[21] R. Cutler, Y. Rui, A. Gupta, J. Cadiz, I. Tashev, L. He, A. Colburn, Z. Zhang, Z. Liu, and S. Silverberg, "Distributed meetings: a meeting capture and broadcasting system," in *Proc. ACM Int. Conf. on Multimedia (MM)*, Juan les Pins, Dec. 2002.

[22] J. DiBiase, *A High-Accuracy, Low-Latency Technique for Talker Localization in Reverberant Environments*, Ph.D. thesis, Brown University, Providence RI, USA, 2000.

[23] A. Doucet, N. de Freitas, and N. Gordon, Eds., *Sequential Monte Carlo Methods in Practice*, Springer-Verlag, 2001.

[24] T.V. Dvorkind and S. Gannot, "Speaker localization using the unscented kalman filter," in *Proc. HSCMA Workshop*, Mar. 2005.

[25] R. Duraiswami, D. Zotkin, and L.S. Davis, "Active speech source localization by a dual coarse-to-fine search," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2001.

[26] D.N. Zotkin and R. Duraiswami, "Accelerated speech source localization via a hierarchical search of steered response power," *IEEE Trans. on Speech and Audio Processing*, vol. 12, no. 5, September 2004.

[27] J. Fisher, T. Darrell, W.T. Freeman, and P. Viola, "Learning joint statistical models for audio-visual fusion and segregation," in *Proc. Neural Information Processing Systems (NIPS)*, Denver, Dec. 2000.

[28] D. Gatica-Perez, G. Lathoud, I. McCowan, J.-M. Odobez, and D. Moore, "Audio-visual speaker tracking with importance particle filters," in *Proc. IEEE Int. Conf. on Image Processing (ICIP)*, Barcelona, Oct. 2003.

[29] 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 Int. Conf. on Computer Vision, Workshop on Multimedia Technologies for E-Learning and Collaboration (ICCV-WOMTEC)*, Nice, Oct. 2003.

[30] D. Gatica-Perez, G. Lathoud, J.-M. Odobez, and I. McCowan, "Multimodal multispeaker probabilistic tracking in meetings," in *Proc. Int. Conf. on Multimodal Interfaces (ICMI)*, Trento, Oct. 2005.

[31] D. Gatica-Perez, G. Lathoud, J.-M. Odobez, and I. McCowan, "Audio-Visual Probabilistic Tracking of Multiple Speakers in Meetings," *IEEE Trans. on Audio, Speech, and Language Processing*, Vol. 15. No. 2. pp. 601-616, Feb. 2007.

[32] T. Gehrig, J. McDonough, "Tracking of Multiple Speakers with Probabilistic Data Association Filters," in *Proc. Workshop on Classification of Events, Actions and Relations (CLEAR)*, Southampton, Apr. 2006.

[33] N. J. Gordon, D. J. Salmond, and A. F. M. Smith, "Novel approach to nonlinear/non-Gaussian bayesian state estimation," in *IEE Proceedings*, 1993, vol. 140, pp. 107–113.

[34] J. Hershey and J. Movellan, "Audio vision: Using audio-visual synchrony to locate sounds," in *Proc. Neural Information Processing Systems (NIPS)*, Denver, Nov. 1999.

[35] M. Isard, *Visual Motion Analysis by Probabilistic Propagation of Conditional Density*, D.Phil. Thesis, Oxford University, 1998.

[36] M. Isard and J. MacCormick, "BRAMBLE: A Bayesian multi-blob tracker," in *Proc. IEEE Int. Conf. on Computer Vision (ICCV)*, Vancouver, Jul. 2001.

[37] S.J. Julier and J.K. Uhlmann, "A new extension of the Kalman filter to nonlinear systems," in *Proc. Int. Sym. on Aerospace/Defense Sensing, Simulation and Controls (AeroSense)*. 1997.

[38] S.J. Julier, J.K. Uhlmann, and H.F. Durrant-Whyte, "A new approach for filtering nonlinear systems," in *Proc. American Control Conf.*, 1995, pp. 1628–1632.

[39] R.E. Kalman, "A new approach to linear filtering and prediction problems," *Trans. of the ASME, Journal of Basic Engineering*, vol. 82, pp. 35–45, March 1960.

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

[41] N. Katsarakis, G. Souretis, F. Talantzis, A. Pnevmatikakis, L. Polymenakos, "3D Audiovisual Person Tracking Using Kalman Filtering and Information Theory," in *Proc. Workshop on Classification of Events, Actions and Relations (CLEAR)*, Southampton, Apr. 2006.

[42] Z. Khan, T. Balch, and F. Dellaert, "An MCMC-based particle filter for tracking multiple interacting targets," in *Proc. European Conf. on Computer Vision (ECCV)*, Prague, May 2004.

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

[44] H. Krim and M. Viberg, "Two decades of array signal processing research: The parametric approach," *IEEE Signal Processing Magazine*, vol. 13, pp. 67 – 94, July 1996.

[45] F. Kubala, S. Colbath, D. Liu, and J. Makhoul, "Rough'n'ready: a meeting recorder and browser," *ACM Computing Surveys*, vol. 31, no. 2es, Jun. 1999.

[46] J.R. Larocque, J.P. Reilly, and W. Ng, "Particle filters for tracking an unknown number of sources," *IEEE Trans. on Signal Processing*, vol. 50, no. 12, December 2002.

[47] G. Lathoud, I.A. McCowan, and J.M. Odobez, "Unsupervised location-based segmentation of multi-party speech," in *Proc. NIST ICASSP Meeting Recognition Workshop*, 2004.

[48] G. Lathoud, J.M. Odobez, and I.A. McCowan, "Short-term spatio-temporal clustering of sporadic and concurrent events," IDIAP-RR 04-14, IDIAP, 2004.

[49] G. Lathoud and I.A. McCowan, "A Sector-Based Approach for Localization of Multiple Speakers with Microphone Arrays," in *Proc. SAPA 2004*, Oct. 2004.

[50] G. Lathoud, J.-M. Odobez, and D. Gatica-Perez, "AV16.3: an Audio-Visual Corpus for Speaker Localization and Tracking," in *Proc. Workshop on Machine Learning for Multimodal Interaction (MLMI)*, Martigny, Jun. 2004.

[51] G. Lathoud and M. Magimai.-Doss, "A sector-based, frequency-domain approach to detection and localization of multiple speakers," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Philadelphia, Mar. 2005.

[52] G. Lathoud, J. Bourgeois, and J. Freudenberger, "Sector-Based Detection for Hands-Free Speech Enhancement in Cars," *EURASIP Journal on Applied Signal Processing, Special Issue on Advances in Multimicrophone Speech Processing*, 2006.

[53] J. LaViola, "A comparison of Unscented and Extended Kalman Filtering for estimating quaternion motion," in *Proc. American Control Conf.*. June 2003, pp. 2435–2440, IEEE Press.

[54] E. Lehmann, *Particle Filtering Methods for Acoustic Source Localisation and Tracking*, Ph.D. thesis, Australian National University, July 2004.

[55] E. Lehmann, "Importance sampling particle filter for robust acoustic source localisation and tracking in reverberant environments," in *Proc. HSCMA Workshop*, Piscataway, NJ, USA, March 2005.

[56] J.S. Liu, *Monte Carlo Strategies in Scientific Computing*, Springer-Verlag, 2001.

[57] M. Lincoln, I. McCowan, J. Vepa, and H.-K. Maganti, "The Multi-Channel Wall Street Journal Audio-Visual Corpus (MC-WSJ-AV): Specifications and Initial Experiments," in *Proc. Automatic Speech Recognition and Understanding Workshop (ASRU)*, San Juan, Dec. 2005.

[58] H.-K. Maganti and D. Gatica-Perez, "Speaker Localization for Microphone-Array-Based ASR: the Effects of Accuracy on Overlapping Speech," in *Proc. Int. Conf. on Multimodal Interfaces (ICMI)*, Banff, Nov. 2006.

[59] H.-K. Maganti, D. Gatica-Perez, and I. McCowan, "Speech Enhancement and Recognition in Meetings with an Audio-Visual Sensor Array," IDIAP Research Report IDIAP-RR-06-24,, submitted to IEEE. Trans. on Audio, Speech, and Language Processing, Apr. 2006

[60] M. Matsumoto and S. Hashimoto, "Multiple signal classification by aggregated microphones," *IEICE Trans. Fundamentals*, vol. E88-A, no. 7, July 2005.

[61] I. McCowan, D. Gatica-Perez, S. Bengio, G. Lathoud, M. Barnard, and D. Zhang, "Automatic analysis of multimodal group actions in meetings," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 27, no. 3, pp. 305–317, Mar. 2005.

[62] 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. on Multimedia (ICME)*, Amsterdam, Jul. 2005.

[63] J.E. McGrath, *Groups: Interaction and Performance*, Prentice-Hall, 1984.

[64] 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 Language Technology Conf. (HLT)*, San Diego, CA, March 2001.

[65] K. Nickel, T. Gehrig, R. Stiefelhagen, and J. McDonough, "A joint particle filter for audio-visual speaker tracking," in *Proc. Int. Conf. on Multimodal Interfaces (ICMI)*, Trento, Oct. 2005.

[66] K. Nickel, T. Gehrig, H.K. Ekenel, J. McDonough, R. Stiefelhagen. "An Audio-visual Particle Filter for Speaker Tracking on the CLEARŠ06 Evaluation Dataset," in *Proc. Workshop on Classification of Events, Actions and Relations (CLEAR)*, Southampton, Apr. 2006.

[67] V. Pavlovic, A. Garg, and J. Rehg, "Multimodal speaker detection using error feedback dynamic bayesian networks," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, Hilton Head Island, SC, 2000.

[68] P. Perez, C. Hue, J. Vermaak, and M. Gangnet, "Color-based Probabilistic Tracking," in *Proc. European Conf. on Computer Vision (ECCV)*, Copenhagen, May 2002.

[69] G.S Pingali, G. Tunali, and I. Carlbom, "Audio-visual tracking for natural interactivity," in *Proc. ACM Int. Conf. on Multimedia (MM)*, Orlando, Oct. 1999.

[70] I. Potamitis, H. Chen, and G. Tremoulis, "Tracking of multiple moving speakers with multiple microphone arrays," *IEEE Trans. on Speech and Audio Processing*, vol. 12, no. 5, Sep. 2004.

[71] S.T. Roweis, "Factorial Models and Refiltering for Speech Separation and Denoising," in *Proc. European Conf. on Speech Communication and Technology (Eurospeech)*, 2003.

[72] Y. Rui, D. Florencio, W. Lam, and J. Su, "Sound source localization for circular arrays of directional microphones," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2005.

[73] R.O. Schmidt, "Multiple Emitter Location and Signal Parameter Estimation," *IEEE Trans. on Antennas and Propagation*, vol. AP-34, pp. 276–280, March 1986.

[74] E. Shriberg, A. Stolcke, and D. Baron, "Observations on overlap: findings and implications for automatic processing of multi-party conversation," in *Proc. European Conf. on Speech Communication and Technology (Eurospeech)*, Aalborg, Sep. 2001.

[75] E. Shriberg, A. Stolcke, and D. Baron, "Can prosody aid the automatic processing of multi-party meetings? Evidence from predicting punctuation and disfluencies, and overlapping speech," in *Proc. ISCA Tutorial and Research Workshop on Prosody in Speech Recognition and Understanding (Prosody)*, 2001.

[76] H. F. Silverman, Ying Yu, J. M. Sachar, and W. R. Patterson III, "Performance of real-time source-location estimators for a large-aperture microphone array," *IEEE Trans. on Speech and Audio Processing*, vol. 13, no. 4, July 2005.

[77] M. Siracusa, L.-P. Morency, K. Wilson, J. Fisher, and T. Darrell, "A multi-modal approach for determining speaker location and focus," in *Proc. Int. Conf. on Multimodal Interfaces (ICMI)*, Vancouver, 2003.

[78] K. Smith, S. Schreiber, I. Potucek, V. Beran, G. Rigoll, D. Gatica-Perez, "2D Multi-Person Tracking: A Comparative Study in AMI Meetings," in *Proc. Workshop on Machine Learning for Multimodal Interaction (MLMI)*, Washington DC, May 2006.

[79] H. Sorenson, *Kalman Filtering: Theory and Application*, IEEE Press, 1985.

[80] R. Stiefelhagen and J. Garofolo (organizers), CLEAR Evaluation Workshop, Southampton, Apr. 2006.

[81] R. Stiefelhagen, J. Yang, and A. Waibel, "Modeling focus of attention for meeting indexing based on multiple cues," *IEEE Trans. on Neural Networks*, vol. 13, no. 4, pp. 928–938, 2002.

[82] D. Sturim, M. Brandstein, and H. Silverman, "Tracking multiple talkers using microphone array measurements," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Munich, Apr. 1997.

[83] J. Vermaak and A. Blake, "Nonlinear filtering for speaker tracking in noisy and reverberant environments," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2001.

[84] J. Vermaak, M. Gagnet, A. Blake, and P. Perez, "Sequential Monte Carlo fusion of sound and vision for speaker tracking," in *Proc. IEEE Int. Conf. on Computer Vision (ICCV)*, Vancouver, July 2001.

[85] B. Vo, S. Singh, and W. K. Ma, "Tracking multiple speakers with random sets," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Montreal, May 2004.

[86] A. Waibel, M. Bett, F. Metze, K. Ries, T. Schaaf, T. Schultz, H. Soltau, H. Yu, and K. Zechner, "Advances in automatic meeting record creation and access," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Salt Lake City, UT, May 2001.

[87] H. Wang and M. Kaveh, "Coherent signal-subspace processing for the detection and estimation of angles of arrival of multiple wide-band sources," *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol. 33, no. 4, August 1985.

[88] D. Ward and R. Williamson, "Particle filter beamforming for acoustic source localization in a reverberant environment," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Orlando, May 2002.

[89] D. Ward, E. Lehmann, and R. Williamson, "Particle filtering algorithms for tracking an acoustic source in a reverberant environment," *IEEE Trans. Speech and Audio Proc.*, vol. 11, no. 6, November 2003.

[90] G. Welch and G. Bishop, "An introduction to the kalman filter," TR 95-041, Dept. of Computer Sc., Uni. of NC at Chapel Hill, 2004.

[91] M. Wolfel, K. Nickel, and J. McDonough, "Microphone Array Driven Speech Recognition: Influence of Localization in the Word Error Rate," in *Proc. Workshop on Machine Learning for Multimodal Interaction (MLMI)*, Edinburgh, Jul. 2005.

[92] B. Wrede and E. Shriberg, "The relationship between dialogue acts and hot spots in meetings," in *Proc. IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, Virgin Islands, Dec. 2003.

[93] T. Yu and Y. Wu, "Collaborative tracking of multiple targets," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, Washington, DC, Jun. 2004.

[94] T. Zhao and R. Nevatia, "Tracking multiple humans in crowded environment," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, Washington DC, Jun. 2004.

[95] D. Zotkin, R. Duraiswami, and L. Davis, "Multimodal 3-D tracking and event detection via the particle filter," in *IEEE Int. Conf. on Computer Vision, Workshop on Detection and Recognition of Events in Video (ICCV-EVENT)*, Vancouver, Jul. 2001.