Multimodal Meeting Monitoring: Improvements on Speaker Tracking and Segmentation through a Modified Mixture Particle Filter

Viktor Rozgić, Carlos Busso, Panayotis G. Georgiou and Shrikanth Narayanan
Department of Electrical Engineering, Speech Analysis and Interpretation Laboratory
University of Southern California, Viterbi School of Engineering
E-mail: rozgic,busso,georgiou,shri@usc.edu

***Abstract*— In this paper we address improvements to our multimodal system for tracking of meeting participants and speaker segmentation with a focus on the microphone array modality. We propose an algorithm that uses *Directions-of-Arrival* estimated for each microphone pair as observations and performs tracking of an unknown number of acoustically-active meeting participants and subsequent speaker segmentation. We propose modified mixture particle filter (mMPF) for tracking of acoustic sources in the *track-before-detection* (TbD) framework. Trajectories of sound sources are reconstructed by the optimal assignment of posterior mixture components produced by mMPF in consecutive frames. Further, we propose a sequential optimal change-point detection algorithm which discovers speech segments in the reconstructed trajectories i.e., performs speaker segmentation. The algorithm is tested on a multi-participant meeting dataset both separately and as a part of the multimodal system. On the task of speaker detection in the multimodal setup we report significant improvement over our previous state of the art implementation.**

I. Introduction

Audio-visual monitoring in multi-participant environments is often used to extract features describing participants' interaction for content annotation. Further processing of the obtained features can provide significant information for content retrieval [1], meeting type classification [2]–[4] and meeting summarization [5].

Our vision is to enable identification and tracking of the dynamics and engagement of participants in meetings [6]. Features of interest for this task are relative positions of meeting participants, speaker identification, and speaker activity and audio event segmentation. In previous work [7] we have presented our smart room system that performs fusion of four different information modalities: a ceiling multi-camera tracking system, a 360° camera face detection system, a microphone array, and a speaker identification system. In this setup the circular microphone array (see Fig. 1) is the only modality that can link the active speaker identity obtained through speaker identification modality to the participant location obtained by tracking from video. In addition to providing complementary information to the other modalities, it also provides redundancy for video localization. Therefore, the overall system performance in identification and localization depends strongly on the quality of the microphone array tracking of sound sources locations and segmentation of speech intervals, the topic of this paper.

Following our previous work [6] we note that the accuracy of the microphone array data fusion method represents a bottleneck for the performance of multimodal tracking of user dynamics. In this work we address this issue by focusing on two problems related to the microphone array modality: tracking of an unknown number of acoustically active participants and active speaker segmentation.

Contribution that we propose is placed in context in Fig.2. Speaker localization, segmentation, and identification all rely heavily on accurate speaker tracking and segmentation of the microphone array output.

We employ a modified *Mixture Particle Filter* (mMPF), based on work by Vermaak et al. [8], to track an unknown number of acoustic sources. The mMPF employs as observations the angular estimates of source locations obtained using the *Fractional Lower Order Statistics Phase Transform* (FLOS-PHAT) method [9] for *Time Difference Of Arrival* (TDOA) estimation for each microphone pair. The nature of the observations is such that it is difficult to design a robust frame level detector of acoustic source appearances and disappearances. For that reason two modifications on the original MPF algorithm are proposed: First, the particle re-clustering step is modified to take into the account both spatial position and weights of particles; and second, MPF is placed within the *Track-before-Detection* (TbD) framework [10] where sources are detected by accumulation of acoustic evidence over time and source trajectories are reconstructed by the optimal two-index [11] assignment of mixture components in consecutive frames. In this formulation the disappearance of acoustic sources is detected when trajectory discontinuity occurs.

In order to discriminate trajectories that belong to active speakers (dominant acoustic sources) from the other acoustic

Fig. 1. On the left is the instrumented conference room and on the right is the 16-microphone array with the omni-directional camera above it.

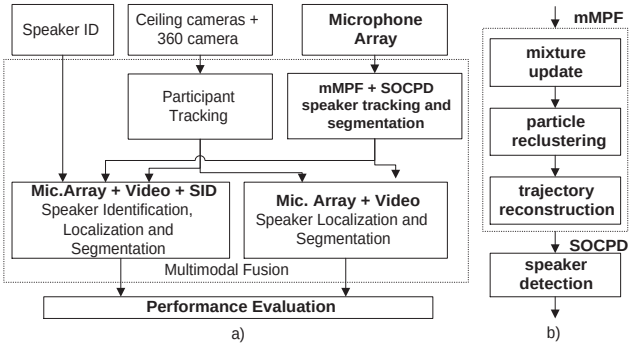

Fig. 2. a) Multimodal system architecture (parts discussed in this work are denoted by bold letters) b) Steps of the proposed microphone array algorithm for speaker tracking and segmentation (proposed algorithm is represented by the mMPF+SOCPD box in multimodal architecture scheme 3a.)

sources (e.g. noise produced by other participants such as paper rattling, coughing etc. as well as sound reflections on surfaces such as the projection screen - see Fig. 1) we apply a *Sequential Optimal Change Point Detection* (SOCPD) algorithm [12] on each reconstructed trajectory. As proposed in Kligys et al. [13], we use separate likelihood statistics for detection of speaker appearances and disappearances and propose a method to compute these statistics from the particle representations of trajectories.

Although MPF interpretation of Vermaak et al. [8], implicitly falls into the TbD category, no particular solution for the trajectory reconstruction was discussed. Other tracking applications of the MPF, such as in [14], do not follow the TbD framework and employ heuristics for detection of appearances and disappearances. For the optimal Bayesian filtering setup, Kligys et al. [13] present a more elaborate treatment of the detection of appearances and disappearances than [15], which proposes an optimal detection method for the particle filtering task. Our method preserves the desirable properties of both frameworks, MPF and TbD, and offers a consistent treatment of trajectory reconstruction and speaker appearnace/disappearance detection.

We test the proposed algorithm in our multi-modal smart room [7]. The proposed 16-microphone, 120-channel data fusion technique, combined with the other modalities brings significant improvement to the overall performance of the smart-room system on speaker tracking and segmentation tasks.

II. Proposed Method

The algorithm we propose can be summarized in four main steps (see Fig. 2b): (i) obtain the posterior distribution of the source locations at each time-frame through the update equations of the mixture particle filter (MPF) (Section II-B); (ii) extract modes of the posterior distribution using the patient rule induction algorithm (PRIM) to (Section II-C); (iii) reconstruct source trajectories by assignment of modes discovered in consecutive time-frames. For this purpose we apply the two index assignment algorithm (Subsection II-C); and (iv) perform speaker segmentation using the *Sequential Optimal Change-Point Detection* (SOCPD) algorithm on each

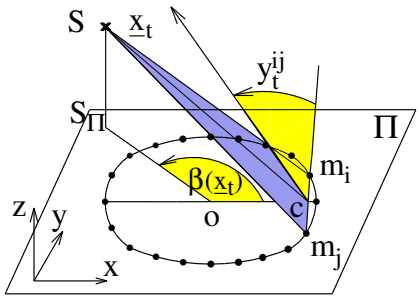

Fig. 3. Microphone array. x_t - position of source S; $\prec (m_i c S)$ - true angle between source S and microphone pair (m_i, m_j); $\beta(x_t)$ - angle that defines source position in the XoY plane; $y_t^{i,j}$ - *Direction of Arrival* estimated by microphone pair (m_i, m_j)

reconstructed trajectory (Subsection II-E).

The statistical model used is described in the following subsection.

A. Statistical Model

We assume that the acoustically active source is represented by its location $\underline{\boldsymbol{x}}_t$ in the quantized 3-D meeting room space.

Our analysis is based on time delay estimates derived by the algorithm described in [9]. Pair-wise delays between M microphones $(m_1, ..., m_M)$ are estimated and transformed to *Direction of Arrival Angles* (DoAA) producing a $M(M-1)/2$-dimensional observation vector $\underline{\boldsymbol{y}}_t$ every time-frame. Given a 16-channel microphone array in our smart room this results in a 120-dimensional observation vector. As shown in Fig. 3 observation coordinate $y_t^{i,j} \in [0, \pi]$ denotes DoAA for the microphone pair (m_i, m_j). Let us denote t-tuple of all observation vectors up to time t as $\underline{\boldsymbol{y}}_{1:t}$.

We assume that the Markovian assumption holds and describe active source kinematics by the transition distribution $p(\underline{\boldsymbol{x}}_t|\underline{\boldsymbol{x}}_{t-1})$ and the initial state distribution $p(\underline{\boldsymbol{x}}_0)$. We compute the observation likelihood $p(\underline{\boldsymbol{y}}_t|\underline{\boldsymbol{x}}_t)$ as:

$$p(\underline{\boldsymbol{y}}_t|\underline{\boldsymbol{x}}_t) = \frac{1}{|\mathcal{R}(\underline{\boldsymbol{x}}_t)|} \sum_{(m_i,m_j)\in\mathcal{R}(\underline{\boldsymbol{x}}_t)} p(y_t^{i,j}|\underline{\boldsymbol{x}}_t), \qquad (1)$$

where $\mathcal{R}(\underline{\boldsymbol{x}}_t)$ denotes the set of all microphone pairs (m_i, m_j) $(i > j)$ for which the distance from the source location $\underline{\boldsymbol{x}}_t$ to the DoAA $y_t^{i,j}$ (see Fig. 3) is smaller than some limiting distance. Both the transition distribution and observation likelihood are learned from a supervised training dataset (see Section III).

Since the goal is to track multiple acoustically active participants, the posterior distribution of interest $p(\underline{\boldsymbol{x}}_t|\underline{\boldsymbol{y}}_{1:t})$ will contain encoded information on the position of each sound source, and hence it is natural to represent it with a mixture model. This approach preserves the low dimensionality of the state space and has clear computational advantages over methods that employ concatenation of the position vectors of the different sources [16]. A computationally efficient approximation of the optimal Bayesian solution is obtained by formulating the tracking problem in the sequential Monte Carlo framework [17], particularly using the mixture particle filter [8] method.

B. Mixture Particle Filter

In this subsection we give a brief overview of the MPF method and describe our choice of the sampling distribution. Vermaak et al. [8] proposed mixture particle filters (MPF) to enable maintaining the multi-modality of the posterior distribution:

$$p(\underline{\boldsymbol{x}}_t|\underline{\boldsymbol{y}}_{1:t}) = \sum_{c=1}^{M_t} \alpha_{c,t} p_c(\underline{\boldsymbol{x}}_t|\underline{\boldsymbol{y}}_{1:t}),$$

where $\alpha_{t,c}$ represents the weight of the c^{th} mixture component at time t. The set $\mathcal{M}_t^c = \{(\underline{\boldsymbol{x}}_{t,c}^i, w_{t,c}^i) : i = 1..N_c\}$ defines a particle approximation of the distribution $p_c(\underline{\boldsymbol{x}}_t|\underline{\boldsymbol{y}}_{1:t})$ where N_c, $\underline{\boldsymbol{x}}_{t,c}^i$ and $w_{t,c}^i$ denote respectively number of particles, position of i^{th} particle and its weight.

MPFs show an elegant way to update particle representation of different mixture components by separate particle filters where the only interaction between different components appears in the particle weight update equations. For more details see Vermaak et al. [8].

A key aspect of all sequential Monte Carlo algorithms is the choice of an appropriate sampling distribution. Particularly, in multi-source tracking scenarios the sampling distribution has to drive particles towards regions where the new sources occur. Therefore, the transition distribution $p(\underline{\boldsymbol{x}}_t|\underline{\boldsymbol{x}}_{t-1})$ does not represent a good choice since it captures only the kinematics of the existing target. In order to overcome this difficulty we use a sampling distribution in the form of linear combination: $p(\underline{\boldsymbol{x}}_t|\underline{\boldsymbol{x}}_{t-1}, \underline{\boldsymbol{y}}_t) = \gamma p(\underline{\boldsymbol{x}}_t|\underline{\boldsymbol{x}}_{t-1}) + (1-\gamma) q(\underline{\boldsymbol{x}}_t|\underline{\boldsymbol{y}}_t)$. Distribution $q(\underline{\boldsymbol{x}}_t|\underline{\boldsymbol{y}}_t)$ is constructed based on agreement between DoAA estimates from different microphone pairs. Note that the triplet $(m_i, m_j, y_t^{i,j})$ defines a conic surface which contains all possible source locations which are indistinguishable from the perspective of the microphone pair (m_i, m_j). For each $\underline{\boldsymbol{x}}_t$ we count how many conic surfaces pass sufficiently close to it and compute distribution $q(\underline{\boldsymbol{x}}_t|\underline{\boldsymbol{y}}_t)$ by normalization of the obtained counts.

C. Particle Reclustering

In order to avoid incorrect identification of particle-mixture pairings we perform particle reclustering. In the MPF algorithm [8] reclustering is performed by a combination of the k-means clustering algorithm and split-merge heuristics. The k-means approach performs clustering based on the positions of particles. This approach suffers since particles are drawn from the sampling distribution and therefore their positions do not follow the true posterior distribution. This problem can not be overcome by resampling on the full particle set because it would undermine preservation of multi-modality as the main principle of MPF.

We propose to solve this problem and determine the number of mixture components without split-merge heuristics by the *Patient Rule Induction Method* (PRIM) [18] briefly described in Table I. Our idea is to use PRIM to detect regions in which the particle approximation has high probability density and adopt these regions as mixture components. This way reclustering is done using both spatial properties of particles and their weights. We define mixture component as a set of particles in the interior of the 3D bounding box with sides parallel to coordinate planes and proceed according to the algorithm in Table I.

TABLE I

BRIEF DESCRIPTION OF THE PRIM ALGORITHM

1. Initalize bounding 3D box with sides parallel to coordinate planes so that it contains all particles
2. Repeat steps 2a and 2b while there is more than N_1 particles in the box.
2a. Cut off ϵ percent of total number of particles in the box by the plane parallel to one of the box's sides. Choose the side in a way that probability density in the remaining part is the biggest possible.
2b. If the new density is smaller than the old one goto 3
3. Repeat steps 3a and 3b while there is less than N_2 particles in the box.
3a. Expand the box along one side so that total number of particles increases for ϵ percent. Choose the side in a way that probability density in the remaining part is the biggest possible.
3b. If the new density is smaller than the old one goto 4.
4. Particles in the obtained box define one mixture component, remove them from the tracking region and repeat steps 1-3 with remaining particles.

D. Trajectory Reconstruction Algorithm

Without the particle reclustering step, MPF performs trajectory maintenance implicitly – a mixture component $\mathcal{M}_t^m$ is obtained by propagation of the particles from a mixture component $\mathcal{M}_{t-1}^m$. The reclustering step interferes with this natural trajectory evolution and redefines mixture components in a way that a component $\mathcal{M}_t^m$ can contain particles obtained by propagation from different components at time $t-1$. Therefore an additional mechanism for trajectory reconstruction is required.

We propose to reconstruct trajectories of acoustic sources by assignment of mixture components in consecutive time-frames. For this purpose we define a metric that describes a similarity between two components. We pose assignment as an optimization problem where the goal is to maximize total similarity between assigned mixture components. This problem can be solved within the integer programming framework. If a mixture component at time t $(t-1)$ is not assigned to a component from $t-1$ (t) we initialize (terminate) a trajectory.

Since particle approximations of mixture components' posterior distributions do not necessarily have identical support sets, it is hard to find a good measure of similarity between them. In order to overcome this problem we fit the Gaussian distribution on each mixture component and use obtained Gaussians to compute inter-component distances.

Lets $s(t, k, m)$ denote symmetrized Kullback-Leibler divergence [19] between Gaussian distributions fitted on mixture components $\mathcal{M}_t^k$ and $\mathcal{M}_{t+1}^m$. We define *goodness of assignment* for these components as:

$$d(t, k, m) = \exp^{-(\alpha_t^k - \alpha_{t+1}^m)^2 - \lambda s(t,k,m)}.$$

The cost of not assigning components is defined as $d(t, 0, m) = d(t, k, 0) = C$. The second term in the exponent favors assignment of components similar in position and shape while the first term favors assignment of components with

similar probability mass. Constants λ and C are determined empirically to fit the application.

Let link variable $c_{k,m}$ for components k and m take value 1 if components are assigned and value 0 if they are not assigned. The optimal assignment is the one that maximizes the total

$$\arg\max_{c_{k,m}} \sum_{k=0}^{M_t} \sum_{m=0}^{M_{t+1}} d(t,k,m) c_{k,m},$$

under the constraint that each mixture component at time t can be assigned to maximally one other component at time $t-1$ and vice versa.

This problem is solved by integer programming technique. Details on how the integer programming algorithms work can be found in Wolsey [11].

E. Detection of Speaker Appearances and Disappearances

Reconstructed trajectories have three possible origins: an active sound source (meeting participant), a temporary fluctuation in the posterior probability caused by reflections or just a reclustering artifact. Our goal is to determine which trajectories belong to meeting participants and segment these trajectories in order to discover intervals that correspond to verbal activity of participants, i.e. to perform speaker segmentation.

For this purpose we propose to apply SOCPD algorithm on each trajectory. A high likelihood that a certain segment of the reconstructed trajectory is produced by a large amount of acoustic evidence (many microphone pairs point in that direction) indicates that such a segment corresponds to the dominant acoustic activity – speech. Further, we conclude that the trajectory on which a speech segment is detected corresponds to a meeting participant. The proposed SOCPD algorithm acts as an additional logic that sequentially discovers start and endpoints of speech segments on reconstructed trajectories. We use separate likelihood statistics for detection of speaker appearances and disappearances (see Klygis et al. [13]) and propose a way to compute these statistics from particle representations of mixture components obtained by the mMPF-TbD tracking algorithm.

Let us assume that the trajectory is represented as a sequence of particle sets $\mathcal{M}_t = \{(\underline{\boldsymbol{x}}_{m,t}, w_{m,t}) : m = 1 \dots N_t\}$ for $t = 1 \dots T_{max}$. Note that for notational simplicity we drop the mixture component indices.

We define a log-likelihood ratio at time t as:

$$l_t := \log \frac{p(\underline{\boldsymbol{y}}_t | \mathcal{M}_{t-1})}{p_0(\underline{\boldsymbol{y}}_t)}. \tag{2}$$

This ratio measures how likely is that observations $\underline{\boldsymbol{y}}_t$ are produced by the sound source at $\mathcal{M}_{t-1}$. Since particles from $\mathcal{M}_{t-1}$ are independent, likelihood $p(\underline{\boldsymbol{y}}_t | \mathcal{M}_{t-1})$ can be computed as:

$$p(\underline{\boldsymbol{y}}_t | \mathcal{M}_{t-1}) = \sum_{m=1}^{N_{t-1}} w_{m,t-1} p(\underline{\boldsymbol{y}}_t | \underline{\boldsymbol{x}}_{m,t-1}) \tag{3}$$

where $p(\underline{\boldsymbol{y}}_t | \underline{\boldsymbol{x}}_{m,t-1})$ is defined by Equation (1). Distribution p_0 represents a uniform distribution on the observation space. Note that we condition on the $\mathcal{M}_{t-1}$ instead of $\mathcal{M}_t$ which is dependent on the observation $\underline{\boldsymbol{y}}_t$. This does not represent a problem in our scenario since the time sampling rate is high enough.

The generalized likelihood ratio AD_{t-1} represents the likelihood that a speaker becomes active at time t_1 and stops his activity at time $t_2 < t$:

$$AD_t := \max_{t_1,t_2 \le t} \log \frac{p(\underline{\boldsymbol{y}}_{t_1:t_2} | \mathcal{M}_{t_1:t_2})}{p_0(\underline{\boldsymbol{y}}_{t_1:t_2})} = \max_{t_1<t_2} \sum_{t=t_1}^{t_2} l_t.$$

The statistic $A_t = \max_{t_1<t} \sum_{\tau=t_1:t} l_\tau$ represents the likelihood that the speaker becomes active at some time $t_1 < t$ and is still active at time t. Therefore statistic $D_t = AD_{t-1} - A_t$ is a measure of the likelihood that a speaker is not active at the time t. The notation used is: D_t - disappeared before time t, A_t - active at time t and AD_t - appeared and disappeared up to time t.

Recursive update rules are given as:

$$AD_t = \max(AD_{t-1}, A_t) = \max(A_t, A_{t-1}, \dots)$$
$$A_t = l_t + \max(0, A_{t-1})$$

where $AD_0 = A_0 = 0$. According to [12] moments of speaker appearance and disappearance can be determined by application of appropriate thresholds on statistics AD_t and D_t respectively.

To summarize, speaker appearance is detected at the moment t_1 at which statistics A_t goes over the first threshold. Speaker disappearance is detected at the time $t_2 > t_1$ at which D_t becomes greater than the second threshold. After a disappearance is detected statistics AD_{t_2} is set to zero and the algorithm is ready to detect a new speech interval.

III. Experimental Results and Discussion

We tested the proposed algorithm on the dataset collected in the University of Southern California smart room [6]. Four sessions with approximate length of 15 minutes each were monitored with multiple modalities: A ceiling 4-camera tracking system, a 360° camera, a single microphone for speaker ID, and a circular 16-microphone array. Microphones were placed in the center of the meeting desk on a ring with 15 cm radius as shown in Fig. 1.

The participants were given multiple topics on which to debate. While they were completely free to follow their beliefs, they were also given a list of arguments to help them along if they needed them. Mostly the interaction ended up being very spontaneous with people seriously believing and arguing for their points of view. This induces frequent changes in the speaker activity i.e., dynamic turn-taking. Monitoring started immediately prior to the people entering the conference room. The average turn duration was 6.727 seconds and in 9.7% of the total speech, different speakers overlapped. More details on the meeting dynamics can be found in Busso et al. [6].

The participants' positions obtained through human annotation from the ceiling multi-camera tracking system are accepted as the reference. Note that the accuracy in geometric space is limited due to the non-point source nature of the human speech production system. The audio data was annotated manually in order to get accurate speaker segmentation. Directions of arrival extracted by processing *Time Difference of Arrival* for each microphone pair are used as observations. We partitioned the dataset in testing (3 sessions) and training (1 session) sets and learned observation likelihoods and transition distributions from the training set.

a) Tracking performance:: In the first experiment we evaluated the tracking performance on intervals on which participant speaks. For this experiment we use mMPF-TbD algorithm. All reconstructed trajectories were analyzed and one closest to the reference trajectory of a participant was assigned to that participant. Average angular error between projections of estimated and true participant's position on the XoY plane (see Fig. 3) on speech intervals was $7.46°$. Note that the nature of observations (120 DoA's) makes it difficult to design a reliable frame level detector of active speaker's position in this scenario. As the relatively low angular errors show, the proposed mMPF-TbD algorithm accumulates evidence through consecutive frames, discovers and maintains tracks for both acoustically dominant and inferior speakers (in 9.7% of total speech time, more than one speaker was active).

b) Speaker segmentation:: In the second experiment we evaluate the performance of the SOCPD algorithm on the speaker segmentation task. Since in our dataset speech represents the most prominent acoustic activity, it was possible to manually determine appropriate threshold values that enable the SOCPD algorithm to recognize speech segments on reconstructed trajectories. In Table II we present statistical properties of the speaker segmentation algorithm which give insights into the behavior of the algorithm in terms of the meeting dynamics.

TABLE II
STATISTICAL PROPERTIES OF THE MMPF-TBD-SOCPD ALGORITHM

avg. duration of speech interval [sec]	6.727
avg. appearance detection delay [sec]	0.421
avg. disappearance detection delay [sec]	0.426
avg. duration of false appearance [sec]	0.545
avg. duration of false disappearance [sec]	0.531
no. of false disapp. per speech interval [overlapping speakers]	0.307
no. of false disapp. per speech interval [non-overlapping]	0.011
avg. duration of non-detected interval [sec]	1.056
total no. of non-detected intervals [overlapping speakers]	45
total no. of non-detected intervals [non-overlapping]	5

Note that 90% (45/50) of non-detected speech intervals take place in segments when multiple participants speak at the same time. Also, average duration of the non-detected speech intervals (1.056sec) is significantly shorter than the overall average (6.727sec), which implies that most speech segments are lost in situations when multiple sources compete for detection. The same holds for false disappearances (non-existing pauses detected within longer speech segments) which are approximately 30 times more likely to occur in intervals where speakers overlap. Values for average delays in detection of start/endpoints of speech intervals as well as the average duration of the falsely detected speech segments are given in the Table II.

c) Multimodal Fusion:: In the third experiment we explore the benefits derived by the proposed algorithm on the performance of our multimodal system [7] on the multimodal speaker segmentation task. We introduce two criteria for judging speaker segmentation quality on 1sec intervals: the strong decision criterion where speaker detection is considered correct if the speaker is active in at least 50% of the 1sec time interval; and the weak decision criterion where speaker detection is considered correct if the speaker is active in any part of the 1sec interval.

Our multimodal system employs a ceiling 4-camera system providing visual hulls of the participants, a $360°$ camera for face tracking, a speaker identification system providing the identities of the current speaker(in this case, equivalent to the seating arrangement), and the 16-microphone array system. In the fusion algorithm (see Fig. 2a) the ceiling cameras and the $360°$ camera system are used to detect number of meeting participants and their locations. In the previous implementation [6] the microphone array system was providing angular position of the active speaker estimated as the mode of the distribution obtained by projecting the directions of arrival for each microphone pair on the XoY plane at each time-frame. In the new implementation we provide estimates of angular active speakers positions in XoY plane obtained by mMPF-TbD algorithm only on intervals in which speakers were actually detected by the SOCPD algorithm. Therefore, the new algorithm introduces two types of improvement: First, on intervals on which multiple speakers were detected it provides multiple angles; and the second, estimates of angular positions of speakers are provided only on intervals on which speakers were actually detected. Speaker detection and localization is performed by probabilistic assignment of angular speakers' positions obtained by the microphone array algorithm to participants locations obtained by video tracking system. Fusion of outputs from microphone array algorithm and the speaker identification system allows multimodal system to learn identities of participants and perform speaker segmentation and localization in parallel. Overview of the multimodal fusion algorithm is presented in Fig. 2a. For more details see [7] and [6].

Performance improvements for both multimodal configurations, *Mic.Array + Video* and *Mic.Array + Video + SID*, and for both performance criteria are presented in Tables III and IV. It is evident that the proposed microphone array algorithm have significant impact on the overall system performance.

Even though performance of the separate speaker identification (SID) system on the speaker detection task (for known assignment of participant identities to spatial locations) is relatively low, 60.10% for strong and 67.85% for the weak detection criteria (see [7]) it provides complementary

TABLE III

PERFORMANCE ON SPEAKER DETECTION TASK: STRONG DECISION

	Old system detection	New system detection	relative gain
Mic. Array+Video	81.97%	83.70%	9.60%
Mic. Array+Video+Speaker ID	83.25%	86.36%	18.57%

TABLE IV

PERFORMANCE ON SPEAKER DETECTION TASK: WEAK DECISION

	Old system detection	New system detection	relative gain
Mic. Array+Video	88.48%	90.22%	15.10%
Mic. Array+Video+Speaker ID	90.57%	93.81%	34.36%

information to the microphone array algorithm and improves overall segmentation performance. This is due to the fact that SOCPD and SID algorithms detect active speaker in different manners: SOCPD does that by monitoring process of competition for observations between different acoustic sources while SID recognizes spectral differences between different speakers and silence. This complementarity adds a new aspect to the multimodal fusion algorithm.

IV. CONCLUSIONS AND FUTURE WORK

In this work we presented improvements in our multimodal system for tracking of meeting participants and speaker segmentation. We achieved these improvements by fusing information obtained by the 16 acoustic channels. We proposed an algorithm that can perform tracking of the acoustically active participants and extraction of speech intervals using *Directions-of-Arrival* estimated for each microphone pair as observations.

Tracking of acoustically active sources was done by use of the modified mixture particle filter (mMPF) in the *Track-before-Detection* (TbD) framework. We modified the original MPF and applied the patient rule induction method (PRIM) to discover mixture components in the posterior distribution. Trajectories were reconstructed by the optimal assignment of discovered mixture components in consecutive time frames. We formulated the optimal mixture component assignment as an integer programming problem and proposed a metric that describes the distance between mixture components. Tracking performance on segments with multiple overlapping speakers shows that mMPF-TbD algorithm can successfully maintain multiple trajectories.

We proposed a novel way to address the speaker segmentation problem by the sequential change-point detection (SOCPD) method. We presented a way to compute statistics used in SOCPD from a particle representation of a reconstructed trajectory. With appropriately tuned threshold values, the SOCPD algorithm applied on particular trajectory discovered time intervals of dominant acoustic activity (speech).

Application of the proposed algorithm in the multimodal setup brought relative speaker detection improvement of 18.57% according to the strong decision criterion and 34.36% according to the weak decision criterion.

The goal for our future research is to augment participant state vector by his/her identity and perform lower level fusion of the observations from the microphone array and speaker ID systems in the SOCPD algorithm by modeling and computing joint likelihoods. Further developments on the tracking side will include analysis of different hierarchical representations of the posterior distribution in the mMPF-TbD and testing the universality of the obtained algorithms on different datasets. Also, we plan to work on the new fusion algorithm which discards inconsistent incoming observations in order to avoid deterioration of the accumulated knowledge on participant identities.

REFERENCES

[1] A. Jaimes, T. Okmura, T. Nagamine, and K. Hirata", "Memory cues for meeting video retrieval," in *In Proc. 1st ACM Workshop of continuous archival and retrieval of personal experiences*, 2004.

[2] I. McCowan, D. Gatica-Perez, S. Bengio, G. Lathoud, M. Barnard, and D.Zhang", "Automatic analysis of multimodal group actions in meetings," *IEEE transactions on on Pattern Analysis and Machine Intelligence*, pp. 305–317, 2005.

[3] S. Reiter, S. Shreiber, and G. Rigoll", "Multimodal meeting analysis by segmentation and classification of meeting events based on higher level semantic approach," in *In Proc. International Conference on Acoustics, Speech and Signal Processing*, march 2005.

[4] S. Banerjee and A. Rudnicky", "Using simple speech based features to detect state of the roles of the meeting participants," in *In Proc. 8th International Conference on Spoken Language Processing*, 2004.

[5] I. Mikić, K. Huang, and M. Trivedi", "Activity monitoring and summarization for an intelligent meeting room," in *In Proc. IEEE Workshop on Human Motion*, december 2000.

[6] C. Busso, P. Georgiou, and S. Narayanan", "Real-time monitoring of participants interaction in a meeting using audio-visual sensors," in *In Proc. International Conference on Acoustics, Speech and Signal Processing*, 2007.

[7] C. Busso, S. Hernanz, C. W. Chu, S. I. Kwon, S. Lee, P. G. Geoargiou, I. Cohen, and S. Narayanan", "Smart room: Participant and speaker localization and identification," in *In Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 2005.

[8] J. Vermaak, A. Doucet, and P. Perez", "Maintaining multimodality through mixture tracking," in *In Proc. Ninth IEEE International Conference on Computer Vision*, october 2003.

[9] P. G. Georgiou, P. Tsakalides, and C. Kyriakakis", "In proc. of the ieee international conference on acoustics, speech, and signal processing," in *Alpha-stable robust modeling of background noise for enhanced sound source localization*, 1999.

[10] P. Williams", "Performance bounds for track-before-detect target detection," in *In Proc.Signal and Data Processing of Small Targets*, 1998.

[11] L. Wolsey", *Integer Programming*. John Wiley & Sons, 1998.

[12] D. Siegmund", *Sequential Analysis*. Springer-Verlag: New York, 1985.

[13] S. Kligys, B.Rozovsky, and A. Tartakovsky", "Detection algorithms and track before detect architecture based on nonlinear filtering for infrared search and track systems," CAMS-University of Southern California, Tech. Rep., 1998.

[14] K. Okuma, A. Taleghani, N. D. Freitas, J. Little, and D. Lowe", "A boosted particle filter: Multitarget detection and tracking," in *In Proc. the European Conference on Computer Vision*, 2004.

[15] Y. Boers and H. Driessen", "A particle filter multi target track before detect application: Some special aspects," in *In Proc. 7th International Conference on Information Fusion*, 2004.

[16] C. Rago, P. Willett, and R. Streit", "A comparison of the jpdaf and pmht tracking," in *In Proc. International Conference on Acoustics, Speech and Signal Processing*, april 1995.

[17] A. Doucet, N. DeFreitas, and N. Gordon", *Sequential Monte Carlo Methods in Practice*. Springer-Verlag: New York, 2001.

[18] T. Hastie, R. Tibshirani, and J. Friedman", *Elements of Statistical Learning*. Springer-Verlag, 2001.

[19] T. M. Cover and J. A. Thomas", *Elements of Information Theory*. John Wiley & Sons, 1991.