

# Multimodal speaker segmentation in presence of overlapped speech segments

Viktor Rozgić, Kyu Jeong Han, Panayiotis 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,kyuhan@usc.edu, georgiou,shri@sipi.usc.edu

## Abstract

We propose a multimodal speaker segmentation algorithm with two main contributions: First, we suggest a hidden Markov model architecture that performs fusion of the three modalities: a multi-camera system for participant localization, a microphone array for speaker localization, and a speaker identification system; Second, we present a novel method for dealing with overlapped speech segments through a likelihood model of the microphone array observations that uses multiple local maxima of the Steered Power Response Generalized Cross Correlation Phase Transform (SPR-GCC-PHAT) function in the Joint Probabilistic Data Association (JPDA) framework. Results show that the proposed method outperforms standard speaker segmentation systems based on: (a) speaker identification and; (b) microphone array processing, for datasets with the significant portion (27.4%) of overlapped speech, and scores as high as 94.4% on the  $F$ -measure scale.

## 1 Introduction

Recently significant research focus has taken place in audio-visual monitoring of multi-participant interactions [1, 2]. Challenging datasets obtained in meeting environments have contributed to the development of many novel signal processing algorithms including multi-target video tracking algorithms which provide relative positions of meeting participants [10], speaker identification (SID) and speaker and audio segmentation algorithms [13]. Outputs of these algorithms have been further used as features in the tasks of automatic content retrieval [12], interaction type classification [15] and content summarization [16].

Our audio-visual smart-room system [6] enables tracking of locations and identities, and performs speaker segmentation, for a variable number of participants. It employs (Figure 1.): (i) a ceiling multi-camera tracking system, (ii) a 360° camera face detection system, (iii) a circular 16-microphone array, and (iv) a SID system. In our previous work we presented contributions on tasks of tracking participant engagement [5] and microphone array processing [20].

Spontaneous interactions usually result in significant speaker overlap, which degrades the quality of automatic speaker segmentation through speaker identification (SID) methods. Recently proposed methods [13, 3] suggest that the speaker segmentation based on microphone array estimation of the direction of arrival (DoA) [4, 7] outperforms



**Figure 1.** Left: instrumented conference room (ceiling camera views); Right: 16-microphone array with the omnidirectional camera above it.

segmentation based on SID techniques [19]. However, the usual monitoring setups in meeting environments include small on-the-desk microphone arrays with limited spatial resolution [6, 10] making it difficult to accurately disambiguate densely-spaced speakers based on DoA cues only. Furthermore, most methods rarely handle overlapped speech at the modeling level and fail to advantageously combine the microphone array and the SID observations.

We suggest a novel design of the speaker segmentation system based on a hidden Markov model (HMM) in which states are the binary speaker activity indicators. The unknown parameter in this HMM is the identity to location mapping, and known parameters are the participants' locations obtained from the video tracking module. This model gives us three levels of flexibility. We can:

- pick the most appropriate method to decode state sequences: Bayesian filtering [8], Viterbi decoding [17], Markov chain Monte Carlo [11].
- use any available likelihood model, based either on microphone array observations only, or on SID observations only, or their combination.
- allow for an easy modification of processing techniques in any modality. Particularly, the system that we proposed in [6] can track unknown number of participants based on combination of the ceiling camera background subtraction and omnidirectional camera face tracking systems. In this work we have opted for a simpler solution, tracking of color markers, in order to focus on the advances in fusion and microphone array processing.

Besides the described specific modality combination (Figure 2.) our main contribution includes a statistical model that enables the microphone array modality to detect multiple overlapping speakers (Section 2.2). For this purpose we

suggest a modification of the *Steered Power Response Generalized Cross-Correlation Phase Transform* (SPR-GCC-PHAT) function [7] in which we re-weight GCC-PHAT functions for different microphone pairs based on their visibility from the different points in the meeting room. Instead of the usual practice where only the global maximum of the SPR-GCC-PHAT function (i.e. the location of the most prominent sound source) is used in sound source localization we suggest extraction of multiple local maxima of the modified SPR-GCC-PHAT function. We treat these maxima as the microphone array observations and use the *Joint Probabilistic Data Association* (JPDA) model [18] to assign them to the active speaker locations. This way we are able to compute the joint likelihood of the microphone array observations when locations of the active speakers are obtained from the video tracking module.

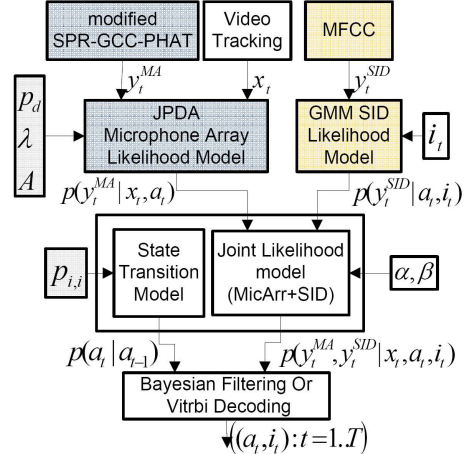
Further, we are able to combine probabilistically (Section 2.4) the microphone array modality and the speaker identification modality. Since our speaker identification modality (Section 2.3) computes likelihoods that a speech frame is produced by one or two concurrent speakers from the known pool of possible participants, we obtain the joint likelihood of all acoustic observations given the locations and identities of the active speakers as a product of microphone array likelihood to the power of  $\alpha$  and the speaker identification likelihood to the power of  $\beta$ . Different choices of the power coefficient pair  $(\alpha, \beta)$  define different fusion models. We argue that the independence assumption is justified by the fact that the microphone array and speaker identification observations are based on different characteristics and processing of the signals.

We tested the automatic segmentation performance using precision and recall measures for sequences of states obtained by Bayesian filtering and Viterbi decoding. In Section 3 we compare performances of baseline likelihood models (SID likelihood model and a recently proposed likelihood model [10] based on the global maxima of the SPR-GCC-PHAT function) with our JPDA likelihood model for microphone array (MA) observations and the proposed combination of the MA and SID likelihood models.

## 2 Proposed method

Our multimodal speaker segmentation algorithm consists of 4 main steps (Figure 2.):

- For the purposes of this publication we track locations of the meeting participants through the detection of color markers and 3D position reconstruction from marker pixel coordinates in multiple views (Section 2.1). Obtained locations are further used as a meta-feature by other modalities.
- We extract local maxima of the modified SPR-GCC-PHAT function and treat them as microphone array observations. Further, we compute likelihood of these observations given positions of active speakers obtained from the first step. We propose a joint probabilistic model for association of microphone array observations to positions of active speakers (Section 2.2).



**Figure 2.** Proposed architecture for multimodal fusion. Parameters  $p_d, \lambda$  and  $A$  are learned from training data. Parameter  $p_{i,i}$  defines state transition and parameters  $(\alpha, \beta)$  define likelihood fusion model.

- We compute likelihoods of speaker identification observations (MFCCs) given speakers' identities. Likelihoods are modeled as Gaussian mixtures for single speakers and overlapped speaker pairs (Section 2.3).
- We decode unknown speaker activity indicators and identity-to-participant associations. We perform fusion of the microphone array and speaker identification likelihoods in the HMM framework for which we define a state transition model. The speaker activity indicator sequence is decoded by both Bayesian filtering and Viterbi algorithm (Section 2.4).

### 2.1 Multi-view color marker tracking

There are many existing techniques in the computer vision community that can be applied for tracking humans in meetingroom environments, e.g. [10, 8]. In our previous work [6] we have presented algorithm that performs such tracking with high accuracy. In order to focus on aspects of microphone array processing and modality fusion, in this publication we track color markers (mini-paper hats) on the participant's heads through 4 ceiling cameras. Different marker colors are described by the Gaussian mixture models in the RGB space and the video system performs detection of the marker pixels and reconstruction of the participants location from multiple views [9].

### 2.2 Microphone array processing

Classical microphone array processing algorithms compute the time domain GCC-PHAT function [4] and estimate the direction of arrival from the global maximum of this function. This way each microphone pair observes only the dominant speaker and it is hard to get correct solution on segments with overlapping speakers. Other solutions, based on the steered power response [7, 10] are used in a similar manner where the global maximum of the SPR-GCC-PHAT function determines location of the dominant speaker.

We propose a modification of the SPR-GCC-PHAT speaker localization algorithm. First, we define a 3D rectangular grid that covers the tracking space; Second, we extract multiple local maxima of the SPR-GCC-PHAT function on the grid and treat their locations as observations. We model association of these observations to the locations of the active speakers (output of the video tracking system) using a joint probabilistic data association model [18]. This model allows active speakers without assigned observations and observations without assigned active speakers.

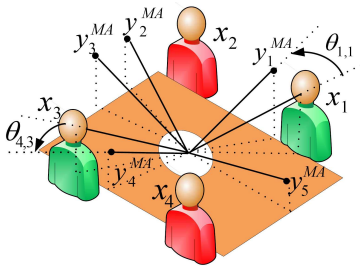
Let us introduce the notation that we use throughout the following sections and describe the modified SPR-GCC-PHAT algorithm in more detail. We assume that  $K_t$  participants are present in frame  $t$  and that their positions obtained from the video tracking system are  $\mathbf{x}_t = (x_{t,1}, \dots, x_{t,K_t})$ . Binary activity indicator vector  $\mathbf{a}_t = (a_{t,1}, \dots, a_{t,K_t}) \in \{0, 1\}^{K_t}$  determines which participants are speaking ( $a_{t,k} = 1$ :  $k^{\text{th}}$  participant is speaking,  $a_{t,k} = 0$ :  $k^{\text{th}}$  participant is silent). Total number of active speakers is  $A_t = \sum_{k=1}^{K_t} a_{t,k}$ .

Observations  $\mathbf{y}_t = (y_{t,1}^{MA}, \dots, y_{t,M_t}^{MA})$  correspond to the  $M_t$  local maxima that are not smaller than  $\gamma \in [0, 1]$  times the value of the global maximum of the modified SPR-GCC-PHAT function given by Equation (1). Parameter  $\gamma$  can be tuned to fit an application.

$$R(y) = \frac{M}{\sum_{\text{all } m} \alpha_m(y)} \sum_{\text{all } m} \alpha_m(y) \mathcal{F}^{-1} \left\{ \frac{S_t^{m_1} S_t^{m_2*}}{|S_t^{m_1} S_t^{m_2*}|} \right\} \quad (1)$$

Functions  $S_t^{m_1}$  and  $S_t^{m_2}$  represent the Fourier transforms of the 100ms Hamming windowed speech segments recorded by the microphone pair  $m = (m_1, m_2)$ .  $M$  is the total number of microphone pairs and  $\mathcal{F}^{-1}$  the inverse Fourier transform operator. We introduce the weighting coefficients  $\alpha_m(y)$  that are equal to one if the location  $y$  is visible by both microphones in the pair  $m$  and is equal to zero otherwise. Additional coefficient  $\frac{M}{\sum_{m=1}^M \alpha_m(y)}$  de-penalizes function value in points on the SPR-GCC-PHAT grid that are not visible from all microphone locations.

The total number of the local maxima  $M_t$  is equal to the sum of number of observations  $M_t^a$  that are corresponding to the active speakers and the number of observations  $M_t^c$  that represent clutter. We assume that  $M_t^c$  follows a Poisson distribution with parameter  $\lambda$ .



**Figure 3.** Sample participant arrangement:  $x_t$  - participants' locations,  $y_t^{MA}$  - local maxima of the SPR-GCC-PHAT function,  $\theta_{i,j}$  - angular distance between  $i^{\text{th}}$  observation and  $j^{\text{th}}$  participant

Further we define the association vector  $\mathbf{r}_t =$

$(r_{t,1}, \dots, r_{t,M_t})$ , where  $r_{t,i} = k$  ( $k = 1, \dots, K_t$ ) for  $a_{t,k} = 1$  means that the observation  $y_{t,i}^{MA}$  is assigned to the active speaker  $k$ . If  $r_{t,i} = 0$  then  $i^{\text{th}}$  observation is not assigned to any speaker. An example of possible data association is given in Figures 3. and 4., where we omit time indices for simplicity.

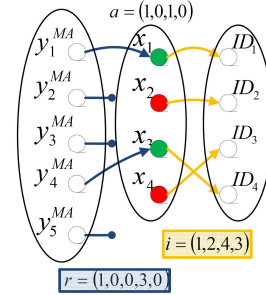
The likelihood of the microphone array observations is given by the Equation (2).

$$p(\mathbf{y}_t^{MA} | \mathbf{a}_t, \mathbf{x}_t) = \sum_{\mathbf{r}_t} p(\mathbf{y}_t^{MA} | \mathbf{a}_t, \mathbf{r}_t, \mathbf{x}_t) p(\mathbf{r}_t | M_t^a, M_t^c) p(M_t^c) \quad (2)$$

If the detection probability for an active speaker is  $p_d$ , term  $p(M_t^a)$  takes value  $\binom{A_t}{M_t^a} p_d^{M_t^a} (1 - p_d)^{A_t - M_t^a}$ . Also, under the reasonable assumption that all valid assignments are equally probable  $p(\mathbf{r}_t | M_t^a, M_t^c) = \left[ \binom{M_t^c}{M_t^c} \cdot K \cdot \dots \cdot (K_t - M_t^a + 1) \right]^{-1}$ . Therefore, the final expression for the microphone array observation likelihood  $p(\mathbf{y}_t^{MA} | \mathbf{a}_t, \mathbf{x}_t)$  is

$$\frac{e^{-\lambda}}{M_t!} \sum_{\mathbf{r}_t} \left( \frac{\lambda}{V} \right)^{M_t^c} p_d^{M_t^a} (1 - p_d)^{A_t - M_t^a} \prod_{\substack{\mathbf{r}_t, i \neq 0 \\ i: a_{t,r_{t,i}} = 1}} p(y_{t,i}^{MA} | x_{t,r_{t,i}}),$$

where  $V$  is the number of SPR-GCC-PHAT grid vertices.



**Figure 4.** Sample associations for the given state:  $p(y^{MA} | a, x, r) = p(\theta_{1,1}) p(\theta_{4,3}) \frac{1}{V^3}$  and  $p(y^{SID} | a, i) = p(y^{SID} | ID_1, ID_4)$

### 2.2.1 Learning of the likelihood model parameters

As it is shown in [10] and [6], in the spherical coordinate system with the origin in the center of the small radius microphone array, source localization techniques based on the TDoA determine angular coordinates of the source much more precisely than its radial coordinate. Therefore, we model the distribution  $p(y_{t,i}^{MA} | x_{t,r_{t,i}})$  as a function of the angular distance (e.g.  $\theta_{1,1}$  and  $\theta_{4,3}$  in Figure 3.) from the observation to the speaker location measured from the center of the microphone array. For the given neighborhood size  $A$  we learn this distribution and other unknown model parameters,  $p_d$  and  $\lambda$ , from the training data.

The probability  $p_d$  is equal to the percentage of participant-frames in which active speakers get at least one observation in neighborhood of size  $A$ . Value  $\lambda$  is the expected number of observations per frame which do not fall in the  $A$ -neighborhood of any active speakers. Finally, the probability distribution of the single observation given a

speaker location is learned as a function of the discretized angular distance from the speaker to the observation where this probability is zero for distances greater than  $A$ .

### 2.3 Speaker identification

Speaker identification systems based on single speaker Gaussian mixture models (GMM) for known set of speakers do not perform well in the presence of the overlapped speech. In order to tackle this difficulty we train GMMs both for single speakers and combinations of two overlapped speakers. For the two-speaker models the corresponding single speaker channels were mixed with equal average energy. Our identification algorithm employs MFCC's on 100ms segments.

Variables  $\mathbf{a}_t$  and  $\mathbf{i}_t$  in combination with the participants' locations obtained from the video tracking module respectively, define locations in space occupied with active speakers, and assign identities of the participants to the particular locations. Therefore, their combination determines identities of the active speakers (Figure 4.) and the speaker identification system can provide the likelihood  $p(\mathbf{y}_t^{SID} | \mathbf{a}_t, \mathbf{i}_t)$ .

### 2.4 Modality fusion and speaker activity decoding

We define a hidden Markov model in which states are the speaker activity indicators  $\mathbf{a}_t$  and in which unknown parameter  $\mathbf{i}_t$  defines the assignment of identities from the known pool of possible speakers to the participant locations  $\mathbf{x}_t$ . In this model we use both the microphone array observations  $\mathbf{y}_t^{MA}$  and the speaker MFCC coefficients  $\mathbf{y}_t^{SID}$  obtained from the SID module.

The short duration of the processing frames makes the synchronous switching of multiple speaker activity indicators very unlikely. Therefore, we allow only those state transitions in which the Hamming distance between consecutive states is less or equal than 1. We define two different state transition models that incorporate this constraint. In the first one we do not allow more than two active speakers per frame while in the second one we pose no limitation on the number of active speakers. The total numbers of states for these models are  $\sum_{i=0}^2 \binom{K_t}{i}$  and  $2^{K_t}$  respectively. For both models we specify a (high) transition probability of staying in the same state  $p(\mathbf{a}_{t+1} = \mathbf{a}_t | \mathbf{a}_t)$  assuming that all other allowed state transitions are equally probable.

Under an independence assumption the joint likelihood can be represented as:

$$p(\mathbf{y}_t^{MA}, \mathbf{y}_t^{SID} | \mathbf{a}_t, \mathbf{x}_t, \mathbf{i}_t) = p(\mathbf{y}_t^{MA} | \mathbf{a}_t, \mathbf{x}_t)^\alpha p(\mathbf{y}_t^{SID} | \mathbf{a}_t, \mathbf{i}_t)^\beta,$$

where different choices of the parameter pair  $(\alpha, \beta) \in [0, 1]^2$  define different likelihood models. We use the following parameter combinations: (i)  $(\alpha, \beta) = (0, 1)$ : speaker identification only; (ii)  $(\alpha, \beta) = (1, 0)$ : microphone array only; (iii)  $(\alpha, \beta) = (1, \beta), \beta \leq 1$ : modality fusion.

Due to the limitation that only models for single and two overlapped speakers are available, the first and the third parameter combination can be used only with the first transition model while the second parameter combination can be used with both state transition models. We found that the

speaker identification likelihood model does not provide reliable disambiguation between states directly connected in the transition model and therefore we discount differences between likelihoods of these states by the parameter  $\beta$ .

The optimal state sequence can be decoded in two ways, as the sequence which maximize posterior state probabilities  $p(\mathbf{a}_t | \mathbf{y}_{1:t}^{MA}, \mathbf{y}_{1:t}^{SID})$  for each  $t = 1, \dots, T$  (optimal Bayesian filtering), or as a sequence that maximizes the posterior probability  $p(\mathbf{a}_{1:T} | \mathbf{y}_{1:T}^{MA}, \mathbf{y}_{1:T}^{SID})$ , where  $T$  is the total number of frames (Viterbi decoding). We apply both techniques with each possible identity-to-location mapping to find the optimal state sequence and mapping.

## 3 Results and discussion

We test the proposed algorithms on two datasets. The first set represents a reading session where four participants read a given text so that their turns significantly overlap. The correct segmentation for this dataset is obtained manually. The second set is a semi-synthetic set obtained, similarly to the [13], by combining and overlapping single speaker segments recorded in the meeting room environment by four different speakers. Total length of the sessions is 15 minutes with 27.4% of the overlapped speech, where the average durations of the segments with one and two active speakers are respectively 8.4s and 3.3s.

Two additional datasets are used to learn parameters of the microphone array likelihood model  $(p_d, \lambda)$  and probability distribution of a single observation given the single speaker location. These sets represent regular meeting sessions with 4 participants, in which the speakers overlap on 8% of the total session length. Model parameter learning is described in more detail at the end of the Section 2.2.

In our experiments not all microphones in the array are visible from all SPR-GCC-PHAT grid points due to the occlusions caused by the omnidirectional camera placed in the center of array (Figure 1.). Therefore, microphone array observations are extracted as the local maxima of the modified SPR-GCC-PHAT function computed in points of the rectangular 20cm grid. For practical purposes, we define the local maxima as the regional maxima in the  $3 \times 3 \times 3$  connected neighborhoods on the grid. Processing is done on 100ms signal segments passed through the Hamming window with 100ms frame shift.

We have computed model parameters for different neighborhood sizes  $A \in \{5, 10, 15, 20, 25, 30, 35, 40\}$  and three different sets of the observations. First set contains all extracted regional maxima; Second set, all regional maxima greater than 55% of the global maximum; and third set, all regional maxima greater than 75% of the global maximum. For all neighborhoods  $A$ , values  $p_d$  and  $\lambda$  rise with increase in number of used observations. Ideally, we want high speaker detection probability  $p_d$  and low probability of false speaker detections. Since the performance with the observation threshold levels 55% and 75% is better than performance with all extracted local maxima we present results only for the second and the third observation set.

Observations for the speaker identification are MFCC coefficients computed on for 100ms frames aligned with mi-

crophone array frames. Gaussian mixture models (GMM) for silence and single participants were trained on 30s training samples, while GMMs for two overlapped participants on training samples obtained by overlapping two single speaker samples with equal average energy.

We evaluated the speaker segmentation performance for 100ms frames using precision ( $P$ ), recall ( $R$ ) and ( $F = \frac{2PR}{P+R}$ ) measures. This type of evaluation is a standard for speaker segmentation type of problems [13, 14] and gives more insight into the performance than the number of correctly detected speaker-frames.

$$P = 100 \frac{\# \text{ of found true active speaker-frames}}{\# \text{ of found active speaker-frames}} \quad (3)$$

$$R = 100 \frac{\# \text{ of found true active speaker-frames}}{\# \text{ of true active speaker-frames}} \quad (4)$$

For the presentation of the experimental results we use the following notation:

- $MA_{0.55}$  and  $MA_{0.75}$  denote microphone array likelihood models,  $(\alpha, \beta) = (1, 0)$ , with observations obtained with threshold levels 55% and 75% respectively.
- $MA_{1.0}$  denotes the baseline microphone array likelihood model presented in the [10]. This method uses only the global maximum of the SPR-GCC-PHAT function as the observation, while the likelihood of this observation given locations of participants is modeled as a product of likelihoods for each participant. Single participant likelihoods take a high constant value when the observation is in the  $A$ -neighborhood of the active speaker or when it is not in the  $A$ -neighborhood of a participant that is not speaking. Otherwise, it takes a low constant value. Both constants and a neighborhood size  $A$  are chosen to maximize  $F$ -measure value.
- $SID$  denotes a likelihood model based on MFCC coefficients extracted in the speaker identification module. We use this model as the second baseline.
- $MA_{xx} \& SID$  denotes the combination of the likelihood models defined in Section 2.4. The parameter pair used for fusion is  $(\alpha, \beta) = (1, 0.5)$ .

We performed exhaustive evaluations of the system performance for different state transition matrices and neighborhood sizes. In this work we present results for the optimal neighborhood size  $A = 10$  and two different state transition models which have in common that not more than one speaker can change activity between two frames and that all allowed state transitions except stay-in-the-same-state,  $p(a_{t+1} = a_t | a_t) = 0.99$ , are equally likely. The only difference is that in one model we introduce the constraint that not more than two speakers can be active in one frame. We present results for two types of HMM state decoding: forward maximum likelihood decoding by the Bayesian filtering and forward-backward optimal sequence decoding by the Viterbi algorithm.

In order to validate the choice  $A = 10$  and  $p(a_{t+1} = a_t | a_t) = 0.99$  which we used throughout the experiments

we present Tables 1. and 2. All values in these tables are obtained for the best likelihood model  $MA_{0.75} \& SID$  for the parameter pair  $(\alpha, \beta) = (1, 0.5)$ . The observed trend is that higher values of the state transition parameter are improving overall segmentation performance, while the neighborhood size  $A = 10$  maximizes performance.

**Table 1.** Performance vs. State Transition Model Parameter  $p(a_{t+1} = a_t | a_t)$  for neighborhood size  $A = 10$

	$p_{i,i}$					
	0.7	0.8	0.9	0.95	0.975	0.99
$P$	94.6	95.0	95.3	95.5	95.7	95.8
$R$	85.6	86.9	88.6	90.6	91.0	93.1
$F$	89.9	90.8	91.8	93.0	93.3	94.4

**Table 2.** Performance vs. Neighborhood Size  $A$  for transition parameter ( $p_{i,i} = 0.99$ )

	$A$							
	5	10	15	20	25	30	35	40
$P$	87.5	95.8	95.9	95.6	95.6	95.8	95.6	95.6
$R$	81.4	93.1	91.2	92.3	92.2	92.3	92.3	92.6
$F$	84.4	94.4	93.6	93.9	93.9	93.9	93.9	94.0

We tested performance of the microphone array likelihood models  $MA_{0.55}$  and  $MA_{0.75}$  that we propose against the baseline model  $MA_{1.00}$  in the setup where we posed no limitations on the number of active speakers. We present these results for both decoding schemes in the Tables 3. and 4. The first three columns represent overall precision ( $P$ ), recall ( $R$ ) and  $F$ -measure; the following three columns contain the same performance measures on segments with one active speaker; and the last three columns contain performance measures on segments with overlapped speech.

**Table 3.** Segmentation Performance: Viterbi Decoding, No Limit on Number of Active Speakers

method	$P$	$R$	$F$	$P_1$	$R_1$	$F_1$	$P_2$	$R_2$	$F_2$
$MA_{1.00}$	98.6	76.5	86.1	98.7	98.2	98.4	99.6	47.7	64.5
$MA_{0.55}$	90.6	88.4	89.5	86.2	88.8	87.5	98.4	87.9	92.9
$MA_{0.75}$	91.5	90.4	91.0	87.3	90.2	88.7	98.8	90.8	94.6

**Table 4.** Segmentation Performance: Bayes Decoding, No Limit on Number of Active Speakers

method	$P$	$R$	$F$	$P_1$	$R_1$	$F_1$	$P_2$	$R_2$	$F_2$
$MA_{1.00}$	96.7	76.7	85.6	96.0	95.7	95.9	100	51.6	68.1
$MA_{0.55}$	85.9	89.0	87.4	79.0	89.8	84.1	98.2	88.0	92.8
$MA_{0.75}$	86.1	90.7	88.4	78.0	91.9	85.3	97.9	89.2	93.3

For the both decoding schemes, our likelihood models ( $MA_{0.55}$  and  $MA_{0.75}$ ) give better overall  $F$ -measure performance than the baseline ( $MA_{1.00}$ ). For Viterbi decoding scheme (Table 3.) on segments with only one active speaker, relative advantage of the baseline over our model is 10.9%. Our model presents a balanced performance both on the segments with a single active speaker (88.7%) and overlapped speech (94.6%). Relative improvement over the baseline on segments with overlapped speech is 46.7%.

Tables 5. and 6. contain results for the transition model that allows at most two active speakers per frame. This transition model allows us to compare all likelihood models.

Like for the transition model that poses no limit on number of active speakers, the proposed MA likelihood models perform better than baselines.

**Table 5.** Segmentation Performance: Viterbi decoding, Maximally 2 Active Speakers per Frame

method	$P$	$R$	$F$	$P_1$	$R_1$	$F_1$	$P_2$	$R_2$	$F_2$
SID	74.7	90.5	81.8	69.8	98.5	81.7	85.1	79.9	82.4
MA <sub>1.00</sub>	98.7	78.1	87.2	98.6	98.8	98.7	100	50.6	67.2
MA <sub>0.55</sub>	93.3	88.8	91.0	90.0	90.7	90.3	99.3	86.4	92.4
MA <sub>0.75</sub>	94.8	90.7	92.7	92.2	92.5	92.4	99.6	88.2	93.6
MA <sub>1.00</sub> & SID	97.7	89.0	93.1	97.7	98.4	98.1	100	77.9	87.6
MA <sub>0.55</sub> & SID	95.5	91.3	93.3	92.8	94.0	93.4	99.9	87.1	93.1
MA <sub>0.75</sub> & SID	95.8	93.1	94.4	93.6	95.2	94.4	99.9	90.3	94.9

**Table 6.** Segmentation Performance: Bayes filtering, Maximally 2 Active Speakers per Frame

method	$P$	$R$	$F$	$P_1$	$R_1$	$F_1$	$P_2$	$R_2$	$F_2$
SID	72.0	87.7	79.1	66.9	95.4	78.7	82.8	77.5	80.1
MA <sub>1.00</sub>	96.0	76.2	84.9	94.9	94.4	94.7	100	52.0	68.4
MA <sub>0.55</sub>	86.9	89.4	88.2	80.5	91.3	85.6	98.7	86.9	92.4
MA <sub>0.75</sub>	87.9	90.8	89.3	82.0	93.2	87.3	98.7	87.5	92.8
MA <sub>1.00</sub> & SID	91.8	88.5	90.1	88.1	97.5	92.5	99.5	76.7	86.6
MA <sub>0.55</sub> & SID	90.5	90.4	90.4	85.6	93.6	89.4	99.5	86.2	92.4
MA <sub>0.75</sub> & SID	91.2	91.1	91.1	87.1	94.7	90.8	99.8	86.4	92.6

The proposed modality fusion model brings further improvement of the performance. This validates our assumption on complementarity of SID and MA likelihood models. Note that the combination of the baseline MA<sub>1.00</sub> and the SID likelihoods degrades MA<sub>1.00</sub> performance on the segments with single active speakers and improves performance on the segments with overlapped speech. On the other side, combination of SID with our models MA<sub>0.55</sub> and MA<sub>0.75</sub> improves performance on all segments for the Viterbi decoding scheme.

## 4 Conclusions and future work

The speaker segmentation system presented in this work is novel from three main perspectives. First, the proposed joint probabilistic data association model (JPDA) uses not only the global maxima of the SPR-GCC-PHAT function as the microphone array (MA) observation, but the multiple regional maxima which allows us to have a better treatment of the situations with speaker overlap. Second, we suggest a hidden Markov model of the speaker activity state evolution which can work with the proposed MA likelihood model only, or perform fusion with the likelihood model obtained from the speaker identification (SID) system. Third, the proposed multimodal architecture performs fusion of the video tracking, MA time delay processing and SID systems and allows for improvements in each modality.

Our JPDA model for the MA observations outperforms the classical speaker segmentation methods, whether these are based on SID [19] or baseline MA based technique [10]. Furthermore, careful thresholding of the extracted regional maxima and the choice of the fusion parameters that emphasizes advantages of both SID and MA likelihood models bring an additional performance improvements.

In our future work we plan to explore influence of the

video tracking accuracy on the segmentation performance, test the proposed algorithm on sessions with dense the participant arrangement and use the segmentation results together with the speaker locations and head orientations obtained from video to analyze and model meeting interaction patterns.

## References

- [1] *Augmented Multiparty Interaction Project*. <http://www.amiproject.org/>.
- [2] *Computers in Human Interaction Loop Project*. <http://chil.server.de/>.
- [3] J. Ajmera, G. Lathoud, and I. McCowan. Clustering and segmenting speakers and their locations in meetings. In *Proc. of the ICASSP*, 2004.
- [4] M. Brandstein and D. Ward. *Microphone Arrays: Signal Processing Techniques and Applications*. Prentice Hall, 2001.
- [5] C. Busso, P. Georgiou, and S. Narayanan. Real-time monitoring of participants interaction in a meeting using audiovisual sensors. In *Proc. ICASSP*, 2007.
- [6] C. Busso, S. Hernanz, C. W. Chu, S. I. Kwon, S. Lee, P. G. Georgiou, I. Cohen, and S. Narayanan. Smart room: Participant and speaker localization and identification. In *Proc. of the ICASSP*, 2005.
- [7] J. Chen, J. Benesty, and Y. Huang. Time delay estimation in room acoustic environments: an overview. *EURASIP Journal on Applied Signal Processing*, 2006.
- [8] A. Doucet, N. DeFreitas, and N. Gordon. *Sequential Monte Carlo Methods in Practice*. Springer-Verlag, 2001.
- [9] D. A. Forsyth and J. Ponce. *Computer Vision, A Modern Approach*. Prentice Hall., 2003.
- [10] D. Gatica-Perez, G. Lathoud, J. M. Odobez, and I. McCowan. Audiovisual probabilistic tracking of multiple speakers in meetings. *IEEE Trans. on Audio, Speech and Language Processing*, 2007.
- [11] P. J. Green. *Nonlinear Dynamics and Statistics*. Birkhauser, 2001.
- [12] A. Jaimes, T. Okmura, T. Nagamine, and K. Hirata. Memory cues for meeting video retrieval. In *Proc. ACM Workshop of archival and retrieval of personal experiences*, 2004.
- [13] G. Lathoud and I. McCowan. Location based speaker segmentation. In *Proc. ICASSP*, 2003.
- [14] J. Makhoul, F. Kubala, R. Schwartz, and R. Weischedel. Performance measures for information extraction. In *Proc. of DARPA Broadcast News Workshop*, 1999.
- [15] 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 PAMI*, 2005.
- [16] I. Mikić, K. Huang, and M. Trivedi. Activity monitoring and summarization for an intelligent meeting room. In *Proc. IEEE Workshop on Human Motion*, 2000.
- [17] L. R. Rabiner. A tutorial on hmm and selected applications in speech recognition. *Proc. of IEEE*, 1989.
- [18] C. Rago, P. Willett, and R. Streit. A comparison of the jpdf and pmht tracking. In *Proc. ICASSP*, 1995.
- [19] D. A. Reynolds and R. C. Rose. Robust text-independent speaker identification using gaussian mixture speaker models. *IEEE Trans. on Speech and Audio Processing*, 1995.
- [20] V. Rozgić, C. Busso, P. G. Georgiou, and S. Narayanan. Speaker tracking and segmentation with microphone array using mixture particle filter: Improvement of multimodal meeting monitoring system. In *Proc. of MMSP*, 2007.