Multimodal speaker segmentation and identification 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 algorithm for speaker segmentation and identification with two main contributions: First, we suggest a hidden Markov model architecture that performs fusion of 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 likelihood model for the microphone array observations for dealing with overlapped speech. We propose a modification of the Steered Power Response Generalized Cross Correlation Phase Transform (SPR-GCC-PHAT) function that takes into the account possible microphone occlusions and use its local maxima as microphone array observations. The Likelihood of the extracted local maxima given positions of active speakers is modeled using the Joint Probabilistic Data Association (JPDA) framework. States in the proposed Hidden markov model are speaker activity indicators and an unknown parameter is mapping of persons’ locations to the set of possible identities. We present and compare two ways for joint estimation of states and the parameter: First, a forward Bayesian filter that performs sequential estimate updates as new observations arrive; Second, a batch optimal decoding using Vitterbi algorithm.

Results show that, for both decoding algorithms, the proposed method outperforms standard speaker segmentation systems based on: (a) speaker identification and; (b) microphone array processing, for the dataset with the significant portion (27.4%) of overlapped speech and scores as high as 94.4% on the F-measure scale.

I. 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 [3], speaker identification (SID) and speaker and audio segmentation algorithms [4]. Outputs of these algorithms have been further used as features in the tasks of automatic content retrieval [5], interaction type classification [6] and content summarization [7].

Our audio-visual smart-room system [8] enables tracking of locations and identities, and performs speaker segmentation, for a variable number of participants. It employs (Fig. 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 [9] and microphone array processing [10].

Spontaneous interactions usually result in significant speaker overlap, which degrades the quality of automatic speaker segmentation through speaker identification (SID) methods. In some situations, e.g. in meetings where participants form sub-groups and start multiple conversations, portion of the overlapped speech is very high and SID methods exhibit poor performance. Recently proposed methods [4], [11] suggest that the speaker segmentation based on microphone array estimation of the direction of arrival (DoA) [12], [13] outperforms segmentation based on SID techniques [14]. However, the usual monitoring setups in meeting environments include small on-the-desk microphone arrays with limited spatial resolution [3], [8] 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 (Fig. 3). The unknown parameter in this HMM is the mapping of participants’ locations to the set of possible identities, 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 [15], Viterbi decoding [16], Markov chain Monte Carlo [17].
- 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 [8] 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 (Fig. 2) our main contribution includes a statistical model that enables the microphone array modality to detect multiple overlapping speakers (Section III-B). For this purpose we suggest a modification of the Steered Power Response Generalized Cross-Correlation Phase Transform (SPR-GCC-PHAT) function [13] 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.

We in order to address speaker overlap by the speaker identification system (SID) we train Gaussian mixture models (GMM) both for single speakers and combinations of two overlapped speakers (by mixing single speaker channels with equal average energy). Speaker activity indicators together with participants’ locations define locations in space occupied with active speakers, which further in combination with identity-to-location mapping provides active speaker identities necessary to compute the SID likelihood.

Further, we are able to combine probabilistically (Section III-D) the microphone array modality and the speaker identification modality. Since our speaker identification modality (Section III-C) computes likelihoods that a speech frame is produced by one or two concurrent speakers from the known pool of possible participants, we model 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.

In Section III-E we propose two methods for estimation of the unknown association parameter and hidden state sequence: First, the sequential Bayesian filtering [15], appropriate for applications where besides performance sequential processing is a demand; and second the Viterbi decoding [16], for applications where data can be post processed and only the performance matters..

We tested the automatic segmentation performance using precision and recall measures [19] for sequences of states obtained by both decoding algorithms. In Section IV present algorithm analysis and compare performances of baseline likelihood models (SID likelihood model and a recently proposed likelihood model [3] 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.

II. PROPOSED METHOD

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

1. We track locations of the meeting participants through the detection of color markers. As mentioned earlier this decorrelates advances in computer vision from audio tracking and fusion developments. We reconstruct the 3D position from marker pixel coordinates in multiple views (Section III-A). The resulting locations are subsequently used as a meta-feature by the other modalities.
2. We extract the local maxima of the modified SPR-GCC-PHAT function and treat them as microphone
array observations. Furthermore, we compute likelihood of these observations given the positions of active speakers as obtained from the first step. We propose a joint probabilistic model for association of microphone array observations to the positions of the active speakers (Section III-B).

3) We compute the likelihoods of speaker identification observations (MFCCs) given the speakers’ identities. Likelihoods are modeled as Gaussian mixtures for single speakers and overlapped speaker pairs (Section III-C).

4) We decode unknown identity-to-participant associations and speaker activity indicators. 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 the Bayesian filtering and the Viterbi algorithms (Section III-D).

III. STATISTICAL MODEL

Let us introduce the notation that we use throughout the following sections. Let $K_t$ be the number of participants present in frame $t$. Their positions are contained in the joint vector $x_t = (x_{t1}, ..., x_{tK_t})^T$, where for each $k$ ($k = 1, ..., K_t$) vector $x_{tk} \in \mathbb{R}^3$ represents location of the $k$th participant in the three dimensional tracking space.

Let $K_{1:t}$ be the total number of different participants’ trajectories registered on interval $[1 : t]$. We define a participant index vector $e_t = (e_{t1}, ..., e_{tK_t})$ in which coordinates are indices of participants present at $t$ in the list of different participant trajectories. Over different frames, all index coordinates assigned to the same participant trajectory obtain the same value from the set $\{1, ..., K_{1:t}\}$ and this value differs from all values assigned to other trajectories.

In the following presentation we will assume that at each time instance $t$ the video tracking system provides us vectors $x_t$ and $e_t$. In other words, we will treat these vectors as known parameters.

The speech activity of the participants present is represented by the binary activity indicator vector $a_t = (a_{t1}, ..., a_{tK_t}) \in \{0, 1\}^{K_t}$, where $a_{tk} = 1$ denotes the $k$th participant is speaking and $a_{tk} = 0$ denotes the $k$th participant is silent. The total number of active speakers at time $t$ is $A_t = \sum_{k=1}^{K_t} a_{tk}$.

Assuming that all participants belong to the finite and known in advance pool of people we define set of possible participants identities as $I = \{1, ..., I\}$. Assuming that the total number of trajectories $K_{1:t}$ registered on the full interval of interest $[1 : T]$ is smaller than $K$ we define the trajectory-to-identity mapping vector $\theta = (\theta_1, ..., \theta_K)$ whose coordinates take values from set $I$. This vector does not evolve through time, and it represents an unknown modeling parameter.

Furthermore, we define a hidden Markov model (see Fig. 3) in which states are the speaker activity indicators $a_t$ and $\theta$ is an unknown parameter. In this model we use both the microphone array observations $y_t^{MA}$ and the speaker ID observations $y_t^{SID}$.

The rapid processing rate, stemming from the short duration of each data frame, 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 model we allow a maximum of two active speakers per frame while in the second one we pose no limitation on the number of active speakers. The total number of states for these models is $\sum_{i=0}^{2} \binom{K_t}{i}$ and $2^{K_t}$ respectively. We specify a (high) transition probability of staying in the same state $p(a_{t+1} = a_t|a_t)$, for both models, while assuming that all other allowed state transitions are equally probable.

Figure 3. HMM: $a_t$ - speaker activity indicator vector; $\theta$ - unknown trajectory-to-identity association parameter

A. Multi-view color marker tracking

There are many existing techniques in the computer vision community that can be applied for tracking humans in meeting room environments, e.g. [3], [15]. In our previous work [8] we have presented a background subtraction algorithm that performs tracking of meeting participants from multiple views with high accuracy. In order to focus on aspects of microphone array processing and modality fusion, we track color markers (mini-paper hats) on the participant’s heads through 4 ceiling cameras. Different marker colors are described by the 3-component Gaussian mixture models in the RGB space and the video system performs detection of the marker pixels and reconstruction of locations $x_t$ of $K_t$ detected participants from multiple views [20]. By keeping track of all registered participant trajectories by the time $t$ we get the vector $e_t$ where its $k$th coordinate is index in the trajectory list.
B. Microphone array likelihood model

Classical microphone array processing algorithms compute the time domain GCC-PHAT function [12] 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 [3], [13] 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.

Observations \( y_t = (y_{1,t}^{MA}, \ldots, y_{M,t}^{MA}) \) 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.

\[
R(y) = \sum_{m}^{M} \alpha_m(y) F^{-1} \left\{ \frac{S_{t}^{m1} S_{t}^{m2*}}{|S_{t}^{m1} S_{t}^{m2*}|} \right\}
\]

Functions \( S_{t}^{m1} \) and \( S_{t}^{m2} \) 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 \( F^{-1} \) 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 \( \frac{M_t}{M} \) 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 assigned active speakers and the number of observations \( M_a \) that represent clutter. We model \( M_t \) as a Poisson random variable with parameter \( \lambda. \)

Further we define the association vector \( r_t = (r_{t,1}, \ldots, r_{t,M_t}), \) where \( r_{t,i} = k(k = 1, \ldots, K_t) \) for \( \alpha_{r,k} = 1 \) means that the observation \( y_{t,k}^{MA} \) is assigned to the active speaker \( k. \) If \( r_{t,i} = 0 \) then \( i \)th observation is not assigned to any speaker. An example of possible data association is given in Figures 4 and 5, where we omit time indices for simplicity.

The likelihood of the microphone array observations can be obtained by averaging over all possible associations and is given by the Equation (2).

\[
p(y_t^{MA}|a_t, x_t) = \sum_{r_t} p(y_t^{MA}|a_t, r_t, x_t) p(r_t|M_t^a, M_t^c) p(M_t^c)
\]
of the angular distance (e.g. $\theta_{1,1}$ and $\theta_{1,3}$ in Figure 4) 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 speaker. 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 $A$.

C. 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 extracted on 100ms segments.

Variables $a_t$ and $\theta$ 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 determine identities of the active speakers (Fig. 5) and the speaker identification system can provide the likelihood $p(y_t^{\text{SID}}|a_t, \theta)$.

D. Modality fusion

Under the assumption that microphone array and speaker identification likelihoods are independent the joint likelihood can be represented as:

$$p(y_t^{\text{MA}}, y_t^{\text{SID}}|a_t, x_t, i_t) = p(y_t^{\text{MA}}|a_t, x_t)^\alpha p(y_t^{\text{SID}}|a_t, i_t)^\beta,$$

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 (maximum two active speakers at a time) while the second parameter combination can be used with both state transition models (no limit on the number of active speakers). 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 $\beta$.

E. Speaker Identity and Activity Decoding

In this section we describe two approaches for decoding of the optimal state sequence and unknown trajectory-to-identity association parameter. For situations where it is necessary to compute estimates of unknown variables sequentially at each time when the new observation arrives, we propose sequential Bayesian filtering approach. In this case the optimal state sequence is defined as:

$$a_1^n, \theta_T = \arg \max_{a_1^n, \theta_T} p(a_1^n, \theta_T|y_{1:T}),$$

and in order to compute it we provide sequential update equations for the state filtering distribution $p(a_t|y_{1:t}, x_t, \theta_T)$ (Section III-E.1).

In situations where we can afford to process the whole session (i.e. interval $[1 : T]$) in a batch, we define the optimal parameter value as $\theta^* = \arg \max_{\theta} p(\theta|y_{1:T}, x_T, \theta_T)$ and the optimal sequence as the path that maximizes joint posterior state probability (Section III-E.2):

$$a_1^n = \arg \max_{a_1^n} p(a_1^n|y_{1:T}, x_t, \theta).$$

In the following sections we simplify the notation by leaving out conditioning on known parameters $x_t$ and $\theta_T$.

1) Sequential Bayesian Filtering: Let’s assume the joint filtering distribution $p(a_{t-1}, \theta|y_{1:t-1})$ at time $t-1$ is known from the previous update step. Then, we compute the predictive distribution as:

$$p(a_t, \theta|y_{1:t-1}) = \sum_{a_{t-1}} p(a_t|a_{t-1})p(a_{t-1}, \theta|y_{1:t-1}).$$

Since there is no observation history at the time $t = 1$, we initialize the update process by the following distribution.

$$p(a_1, \theta) = \begin{cases} \prod_{k=1}^{K} p(a_{k,1}|0) & \text{if } \theta - \text{allowed assignment} \\
0 & \text{otherwise} \end{cases}.$$

Number of the possible trajectory-to-identity associations is $|\Theta| = I(I-1)\ldots(I-K+1)$ and size of the joint state-parameter space is $2^{K_1}|\Theta|$

We obtain the state filtering distribution by marginalization over the space of the unknown parameter $\theta$, i.e. by averaging over all trajectory-to-identity associations:

$$p(a_t|y_{1:t}) \sim \sum_{\theta} p(y_{t}|a_t, \theta)p(a_t, \theta|y_{1:t-1}).$$

After the new observation $y_t$ is introduced we get the updated distribution of unknown parameter $p(\theta|y_{1:t})$ by marginalization over the state space:

$$p(\theta|y_{1:t}) \sim \sum_{a_t} p(y_{t}|a_t, \theta)p(a_t, \theta|y_{1:t-1}).$$

Finally, since until the time $t$ only $K_{1:t}$ trajectories were registered, the meaningful part of the parameter distribution is $p(\theta_{1:K_{1:t}}|y_{1:t})$ and we obtain it by marginalization over $\theta_{K_{1:t}+1:K_t}$. 

We apply all mentioned steps for the estimation of the state and the parameter filtering distributions sequentially for each $t = 1, \ldots, T$ which we summarize in the Algorithm 1.

**Algorithm 1 Sequential Bayesian Filtering**

```plaintext
for $t = 1$ to $T$ do
    (A) Compute $p(a_t, \theta | y_{1:t-1})$:
    if $t = 1$ then
        Use Equation 7.
    else
        Use Equation 6.
    end if
    (B) Compute $p(a_t | y_{1:t})$. Use Equation 8.
    (C1) Compute $p(\theta | y_{1:t})$. Use Equation 9.
    (C2) Compute $p(\theta | y_{1:t}, \lambda) = \sum_{\theta_{K_1+1,K}} p(\theta | y_{1:t})$
end for
```

Total complexity of the forward-sequential filtering algorithm on interval $[1 : T]$ is $O(T \cdot 2^K \cdot |\Theta|)$.

2) **Viterbi Decoding**: In this algorithm we use the same update equations to get distribution of the unknown association parameters. Only, at the final time $T$ complete information on existing trajectories and their overlap in time is available. Therefore, at $T$ it is possible to prune the set of possible trajectory-to-identity associations $\Theta$. This decreases the overall computational complexity.

**Algorithm 2 Viterbi Decoding**

```plaintext
for $t = 1$ to $T$ do
    (A) Compute $p(a_t, \theta | y_{1:t-1})$:
    if $t = 1$ then
        Use Equation 7.
    else
        Use Equation 6.
    end if
    (B) Compute $p(\theta | y_{1:t})$. Use Equation 9.
end for

(C) Find optimal parameter $\theta^*$. Use Equation 5.
(D) Find optimal state sequence $a_{1:T}$. Use Equation 5.
```

Steps (C) and (D) have complexity $O(T \cdot 2^K)$ and therefore, total complexity of the Viterbi decoding algorithm is effectively the same as for the sequential Bayesian filter.

**IV. 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 [4], 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 III-B.

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 (Fig. 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.

Probability $p_{fd}$ presented in Table I, although not a modeling parameter, represents percentage of participant-frames in which a non-active speaker gets at least one observation in the $A$-neighborhood and gives insight in the dependency of the false speaker detections on number of extracted local maxima. Table I contains the model parameters computed for different values of speaker neighborhood $A$. Parameters in the row $R_{LM} > 0$ are computed using all regional maxima while the following two rows $R_{LM} > 0.55R_{GM}$ and $R_{LM} > 0.75R_{GM}$ correspond respectively to the cases where only the regional maxima higher than 55% and 75% of the global maxima $R_{GM}$ are used as observations. Values $p_d$ and $p_d$ rise with relaxation of the neighborhood size $A$ while the expected number of false alarms $A$ falls. Ideally, we want high speaker detection probability $p_d$ and low probability of false detections $p_{fd}$. Value $p_{fd}$ (Table I) rises significantly with the number of extracted local maxima. Having this in mind, we conduct all experiments for the threshold levels
55% and 75%.

<table>
<thead>
<tr>
<th>Neighborhood Size</th>
<th>Performance Measures</th>
<th>Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>Precision (P)</td>
<td>94.6</td>
</tr>
<tr>
<td>10</td>
<td>Recall (R)</td>
<td>95.0</td>
</tr>
<tr>
<td></td>
<td>F1-Score (F)</td>
<td>95.3</td>
</tr>
<tr>
<td></td>
<td></td>
<td>95.5</td>
</tr>
<tr>
<td></td>
<td></td>
<td>95.7</td>
</tr>
<tr>
<td></td>
<td></td>
<td>95.8</td>
</tr>
</tbody>
</table>

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

- **MA** denotes microphone array like

hood size gives more insight into the performance than the number

of speaker segmentation type of problems [4], [19] and

100

for two single speaker samples with equal average energy.

Participants on training samples obtained by overlapping

30

(GMM) for silence and single participants were trained

with microphone array frames. Gaussian mixture models

pr

falls. Ideally, we want high speaker detection probability

p

and low probability of false detections pfd. Value

pfd (Table I) rises significantly with the number of

extracted local maxima. Having this in mind, we conduct

all experiments for the threshold levels 55% and 75%.

Observations for the speaker identification are MFCC

coefficients computed on for 100ms frames aligned

with microphone 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 =

2PR/P + R). This type of evaluation is a standard

for speaker segmentation type of problems [4], [19] and

gives more insight into the performance than the number

of correctly detected speaker-frames.

P = 100 \# of found true active speaker-frames |

\# of found active speaker-frames

(10)

R = 100 \# of found true active speaker-frames |

\# of true active speaker-frames

(11)

For the presentation of the experimental results we use the

following notation:

- **MA** and **MA** denote microphone array like

hood models, \((\alpha, \beta) = (1, 0)\), with observations

obtained with threshold levels 55% and 75% respec

vatively.

- **MA** denotes the baseline microphone array like

hood model presented in [3]. 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** and **SID** denotes the combination of the like

hood models defined in Section III-D. 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 sequential 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 II and III. All values in these tables are

obtained for the best likelihood model MA and 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 II.

<table>
<thead>
<tr>
<th>Neighborhood Size</th>
<th>Precision (P)</th>
<th>Recall (R)</th>
<th>F1-Score (F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>94.6</td>
<td>95.0</td>
<td>95.3</td>
</tr>
<tr>
<td>15</td>
<td>95.6</td>
<td>96.9</td>
<td>97.3</td>
</tr>
<tr>
<td>20</td>
<td>95.6</td>
<td>96.9</td>
<td>97.3</td>
</tr>
<tr>
<td>30</td>
<td>95.6</td>
<td>96.9</td>
<td>97.3</td>
</tr>
<tr>
<td>40</td>
<td>95.6</td>
<td>96.9</td>
<td>97.3</td>
</tr>
</tbody>
</table>

### TABLE III.

<table>
<thead>
<tr>
<th>Neighborhood Size</th>
<th>Precision (P)</th>
<th>Recall (R)</th>
<th>F1-Score (F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>87.5</td>
<td>95.8</td>
<td>95.6</td>
</tr>
<tr>
<td>10</td>
<td>88.4</td>
<td>94.4</td>
<td>93.9</td>
</tr>
</tbody>
</table>
columns contain performance measures on segments with overlapped speech.

### TABLE IV. SEGMENTATION PERFORMANCE: VITERBI DECODING, NO LIMIT ON NUMBER OF ACTIVE SPEAKERS

<table>
<thead>
<tr>
<th>method</th>
<th>P</th>
<th>R</th>
<th>F</th>
<th>P1</th>
<th>R1</th>
<th>F1</th>
<th>P2</th>
<th>R2</th>
<th>F2</th>
<th>P3</th>
<th>R3</th>
<th>F3</th>
<th>P4</th>
<th>R4</th>
<th>F4</th>
</tr>
</thead>
<tbody>
<tr>
<td>MA1.00</td>
<td>98.6</td>
<td>86.5</td>
<td>86.1</td>
<td>98.2</td>
<td>98.3</td>
<td>98.4</td>
<td>99.6</td>
<td>47.7</td>
<td>64.8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MA0.55</td>
<td>91.6</td>
<td>88.4</td>
<td>89.2</td>
<td>88.8</td>
<td>87.3</td>
<td>88.3</td>
<td>87.9</td>
<td>92.9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MA0.75</td>
<td>91.5</td>
<td>90.4</td>
<td>91.0</td>
<td>87.3</td>
<td>90.2</td>
<td>88.7</td>
<td>98.8</td>
<td>90.8</td>
<td>94.6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Similarly to the transition model that poses no limit on number of active speakers, the proposed MA likelihood models perform better than the baseline ones.

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 MA1.00 and the SID likelihoods degrades MA1.00 performance on the segments with single active speaker and improves performance on the segments with overlapped speech. On the other side, combination of SID with our models MA0.55 and MA0.75 improves performance on all segments for the Viterbi decoding scheme.

### V. CONCLUSIONS

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 better handling in regions of speaker overlap. Our JPDA model for the MA observations outperforms the classical speaker segmentation methods on segments with overlapped speech, whether these are based on SID [14] or baseline MA based technique [3]. Furthermore, careful thresholding of the extracted regional maxima and the choice of the fusion parameters that emphasize advantages of both SID and MA likelihood models bring additional performance improvements.

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. This multimodal architecture performs fusion of the video tracking, MA time delay processing and SID systems and allows for improvements in each modality.

Finally, we propose two probabilistic algorithms that solve the interesting problem of parallel estimation of the unknown trajectory-to-identity association parameter and state sequence.

### VI. APPENDIX

In order to validate our proposed likelihood model for the microphone array observations we compare it with the recently proposed model [3] that uses only the global maximum of the SPR-GCC-PHAT function. For completeness we briefly present here this baseline model.

For the observations $y_t = (y_{t,1}, \ldots, y_{t,M_A})$, locations of the participants $x_t = (x_{t,1}, \ldots, x_{t,K_A})$ and the speaker activity vector $a_t = (a_{t,1}, \ldots, a_{t,K_A})$ the joint likelihood is defined as:

$$p(y_t|x_t, a_t) = \prod_{i:a_{t,i}=1} p_1(y_{t,i}|x_{t,i}) \prod_{i:a_{t,i}=0} p_0(y_{t,i}|x_{t,i}),$$
where the probabilities $p_1$ and $p_0$ are given by the following equations:

$$p_1(y_t|x_{t,i}) = \begin{cases} L_1(j) : \|y_{t,j} - x_{t,i}\| \leq A \\
L_0(j) : \|y_{t,j} - x_{t,i}\| \geq A \end{cases}$$

$$p_0(y_t|x_{t,i}) = \begin{cases} L_1(j) : \|y_{t,j} - x_{t,i}\| \geq A \\
L_0(j) : \|y_{t,j} - x_{t,i}\| \leq A \end{cases}$$

The constant $A$ defines a neighborhood size and the constants $L_1$ and $L_0$ are chosen to favor observation existence for each active speaker, and observation absence for all participants that are not speaking. Therefore ratio $\frac{L_1}{L_0}$ has to be significantly greater than one.

**ACKNOWLEDGMENT**

**REFERENCES**