Multimodal speaker segmentation and identification in presence of overlapped speech segments

Viktor Rozgic, Kyu J. Han, Panayiotis G. Georgiou and Shrikanth Narayanan
Department of Electrical Engineering, Signal 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 describe a multimodal algorithm for speaker segmentation and identification with two main contributions: First, we propose a hidden Markov model architecture that performs fusion of three information sources: 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 account the 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.

The state in the proposed hidden Markov model is a vector of the speaker activity indicators of present participants, and the unknown parameter is the mapping of participants’ locations to the set of all possible participants’ identities. We present and compare two ways for the joint estimation of the states and the unknown parameter: the first, a forward Bayesian filter that performs sequential estimate updates as new observations arrive and the second, a batch decoding using the Viterbi 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 dataset with significant portion (27.4%) of overlapped speech and scores as high as 94.4% on the F-measure scale.

Index Terms—multimodal fusion, speaker segmentation, microphone array, joint probabilistic data association, Bayesian filtering, parameter estimation

I. INTRODUCTION

An audio-visual monitoring of multi-participant interactions is an emerging area of research in multimedia signal processing [1], [2]. Challenging datasets obtained in meeting environments have contributed to the development of many novel signal processing algorithms including multi-target video tracking 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 speaker 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]. In this work we tackle the problem of robustly processing overlapping speech with a multimodal approach.

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

Spontaneous interactions usually result in significant speaker overlap, which degrades the quality of automatic speaker segmentation through speaker identification (SID) methods [11]–[15]. In some situations, e.g. in meetings where participants form sub-groups and start multiple conversations, the portions of the overlapped speech are very high and SID methods exhibit poor performance. Recently proposed methods [4], [16] suggest that the speaker segmentation based on microphone array estimation of the direction of arrival (DoA) [17], [18] outperforms segmentation based on SID techniques [19]. However, the typical 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 methods.
We suggest a novel design of the speaker segmentation and identification system based on the hidden Markov model (HMM) (Fig. 3) in which state is obtained by concatenation of the binary speaker activity indicators for all present participants. The unknown parameter in this HMM is the mapping of participants’ locations to elements of the set of all possible participant identities. The known parameters are the participant 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 [20], [21], Viterbi decoding [22], Markov chain Monte Carlo [23], [24],
- 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 using a 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.

In addition to 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). Towards this goal we suggest a modification of the Steered Power Response Generalized Cross-Correlation Phase Transform (SPR-GCC-PHAT) function [18] 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 [25] to assign them to the active speaker locations. This allows computation of joint likelihood of the microphone array observations when locations of the active speakers are obtained from the video tracking module.

While different methods for speaker segmentation are available [26]–[30] in order to incorporate SID in the multi-modal probabilistic model for speaker segmentation we choose modification of the SID method presented in [13], [14], [19]. In order to address the speaker overlap by the speaker identification (SID) system we train Gaussian mixture models (GMM) for every single speaker and all combinations of two overlapped speakers. We obtain training data for the models of overlapped speech by mixing single speaker channels with equal average energies. Speaker activity indicators together with participants’ locations define locations in space occupied with active speakers, which further in combination with location-to-identity mapping provides active speaker identities necessary to compute the SID likelihoods.

Furthermore we are able to combine probabilistically (Section III-D) the microphone array modality and the speaker identification modality. The latter provides likelihoods that a speech frame is produced by one or two concurrent speakers from the known pool of possible participants. By introduction of reasonable simplifying assumptions that microphone array observations (directions-of-arrival) do not depend on active speakers’ identities and speaker ID observations (MFCC coefficients), and that the speaker ID observations do not depend on speakers’ locations we are able to model the joint likelihood of all acoustic observations given the locations and identities of 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 power coefficient pairs $(\alpha, \beta)$ define different fusion models.

We propose (Section III-E) two methods for the estimation of the unknown association parameter and the hidden state sequence. The first method, sequential Bayesian filtering [20], [21], is appropriate for applications where both performance and sequential processing are important. The second method, Viterbi decoding [22], should be the method of choice for applications where only the performance is important and data can be post processed.

We tested the automatic segmentation performance using precision and recall measures [31] for sequences of states obtained by both decoding algorithms. In Section IV we 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 local maxima of the modified SPR-GCC-PHAT function and treat them as microphone array observations. Furthermore, 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 III-B).
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 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 Bayesian filtering and Viterbi algorithm (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 \( x_t = (x_{t,1}^T, \ldots, x_{t,K_t}^T)^T \), where for each \( k \) \((k = 1, \ldots, K_t)\) vector \( x_{t,k}^T \in \mathbb{R}^3 \) represents location of the \( k^{th} \) participant in the three dimensional tracking space.

Let \( L_{1:t} \) be the total number of different participants’ trajectories registered on the interval \([1, t]\). We define the trajectory index vector \( e_t = (e_{t,1}, \ldots, e_{t,K_t}) \), such that the trajectory of the \( k^{th} \) participant at frame \( t \) has the index \( e_{t,k} \) in the list of all trajectories detected on the interval \([1, t]\). All trajectory index coordinates assigned to the same participant’s trajectory obtain the same value from the set \([1, \ldots, L_{1:t}]\) and this value differs from the values assigned to other trajectories.

In the following presentation we assume that at each frame \( t \) the video tracking system provides vectors \( x_t \) and \( e_t \) (i.e. we consider the position vector \( x_t \) and trajectory index vector \( e_t \) as a known parameters).

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

Assume that all participants belong to the finite pool of people whose identities are known in advance. Let us denote set of these identities as \( I = \{1, \ldots, I\} \). Assuming that the total number of trajectories \( L_{1:t} \) registered on the full interval of interest \([1, T]\) is smaller than \( L \), we define the trajectory-to-identity mapping vector \( \theta = (\theta_1, \ldots, \theta_L) \) whose coordinates take values from \( I \). 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. As mentioned earlier vectors \( x_t \) and \( e_t \) are known parameters. 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 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 \( \sum_{i=0}^{2} (K_t) \) 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 2. Proposed architecture for multimodal fusion. Parameters \( p_d, \lambda \) and \( \lambda \) are learned from training data. Parameter \( p_{i,\lambda} \) defines state transition and parameters \( (\alpha, \beta) \) define likelihood fusion model.

### 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], [20]. In our previous work [8] we 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 [32]. By updating the list of all registered participant trajectories on the interval $[1, t]$ we get the vector $e_t$ where its $k^{th}$ coordinate $e_{t,k}$ is an index in the trajectory list.

**B. Microphone array likelihood model**

Classical microphone array processing algorithms compute the time domain GCC-PHAT function [17] 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 difficult to get correct solution on segments with overlapping speakers. Other solutions, based on the steered power response [3], [18] 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 [25]. 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.

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

(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 $m = (m_1, m_2)$. Product ‘\cdot’ denotes component by component multiplication, while || denotes $L_2$ norm. $M$ 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 $\frac{M}{\sum_{m} \alpha_m(y)}$ de-penalizes the 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 $M_t^r$ that are corresponding to the active speakers and the number of observations $M_t^c$ that represent clutter. We model $M_t^c$ as a Poisson random variable with parameter $\lambda$. Use of the Poisson distribution is appropriate in situations where probability of the event (detection of local maxima) is uniform in space. The process of local maxima extraction captures local fluctuations in the GCC-PHAT function. In regions without prominent maxima, i.e. actual speaker locations, it gives relatively uniform spatial distribution of local maxima. Therefore, we decided to adopt the usual choice from the data association literature and use the Poisson distribution.

![Sample participant arrangement](image)

We relate the observation $y_{t,i}^{MA}$ to the active speaker $k$ through an association vector $r_t = (r_{t,1}, \ldots, r_{t,M_t})$ where $r_{t,i} = k$ for $i = 1, \ldots, K_t$ and $r_{t,i} = 0$ if $i > K_t$. Observation $y_{t,i}^{MA}$ 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 Fig. 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 assignments and is given by the Equation (2).

$$p(y_t^{MA}|a_t, x_t) = \sum_{r_t} p(y_t^{MA}|a_t, r_t, x_t)p(r_t|M_t^a, M_t^c)p(M_t^c)$$

(2)

If the detection probability for an active speaker is $p_d$ then:

$$p(M_t^a) = \left( \frac{A_t}{M_t^a} \right)^{M_t^a} \left( 1 - p_d \right)^{A_t - M_t^a}. $$

Also, under the assumption that all valid assignments are equally probable:

$$p(r_t|M_t^a, M_t^c) = \left( \frac{M_t^a}{M_t^a} \right) \cdot K \cdot \ldots \cdot (K_t - M_t^a + 1)^{-1}. $$

© 2010 ACADEMY PUBLISHER
Likelihood of the microphone array observations given active speaker locations and observation-to-speaker association, under the assumption that clutter has uniform spatial distribution, can be factorized in terms corresponding to different observations and written as:

\[
p(y_t^{MA}|a_t, x_t) = \left( \frac{1}{V} \right)^{M_t^a} \prod_{i, r_t \neq 0} p(y_t^{MA}|x_{t,r_t}),
\]

where \( V \) is the number of SPR-GCC-PHAT grid vertices.

Therefore, after substitution of all product terms in Equation 2 the final expression for the microphone array observation likelihood \( p(y_t^{MA}|a_t, x_t) \) becomes:

\[
e^{-\lambda} \sum_{r_t} \left( \frac{1}{V} \right)^{M_t^a} p_d^{M_t^a} (1-p_d)^{M_t^a} \prod_{i, r_t \neq 0} p(y_t^{MA}|x_{t,r_t}),
\]

C. Speaker identification

Speaker identification systems based on single speaker Gaussian mixture models (GMM) for a known set of speakers do not perform well in the presence of overlapped speech. In order to tackle this difficulty we train 16-component GMMs with diagonal covariance matrices 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. In order to have the same analysis window as for the microphone array processing our speaker identification algorithm employs MFCCs extracted on 100ms segments. For the the chosen window length (100ms) we decided not to use delta and delta-delta features.

Variables \( a_t \) and \( \theta \) in combination with the participant 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 (Fig. 5) and the speaker identification system can provide the likelihood \( p(y_t^{SID}|a_t, \theta) \).

D. Modality fusion

Without loss of generality we can factorize the joint likelihood as:

\[
p(y_t^{MA,SID}|a_t, x_t) = p(y_t^{MA}|y_t^{SID}, a_t, x_t) p(y_t^{SID}|a_t, \theta, x_t),
\]

where different parameter pairs \((\alpha, \beta) \in [0, 1]^2\) define different likelihood models. Also, under reasonable simplifying assumptions that \( p(y_t^{MA}|y_t^{SID}, a_t, x_t) = p(y_t^{MA}|a_t, x_t) \) and \( p(y_t^{SID}|a_t, \theta, x_t) = p(y_t^{SID}|a_t, \theta) \), the joint likelihood model becomes a combination of the described microphone array and speaker identification likelihood models:

\[
p(y_t^{MA,SID}|a_t, x_t) = p(y_t^{MA}|a_t, x_t) p(y_t^{SID}|a_t, \theta),
\]

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) \((\alpha, \beta) = (1,1), \beta \leq 1\): 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 (maximally one speaker activity change per frame) 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 the unknown variables sequentially at each time when a new observation arrives, we propose the sequential Bayesian filtering approach. In this case the optimal state sequence \( \mathbf{a}_{1:T}^* \) is defined as:

\[
\mathbf{a}_{1:T}^* = \left( \arg \max_{\mathbf{a}_t} p(\mathbf{a}_t | \mathbf{y}_{1:T}^{MA,SID}, \mathbf{x}_t, \mathbf{e}_t) \right)_{1:T},
\]

and in order to compute it we provide sequential update equations for the state filtering distribution \( p(\mathbf{a}_t | \mathbf{y}_{1:T}^{MA,SID}, \mathbf{x}_t, \mathbf{e}_t) \) (Section III-E.1).

In situations where we can afford to process the whole session, \([1, T]\), in a batch, we define the optimal parameter value as \( \theta^* = \arg \max_{\theta} p(\theta | \mathbf{y}_{1:T}^{MA,SID}, \mathbf{x}_1:T, \mathbf{e}_1:T) \) and the optimal sequence as the path that maximizes joint posterior state probability (Section III-E.2):

\[
\mathbf{a}_{1:T}^* = \arg \max_{\mathbf{a}_{1:T}} p(\mathbf{a}_{1:T} | \mathbf{y}_{1:T}^{MA,SID}, \mathbf{x}_1:T, \mathbf{e}_1:T, \theta^*).
\]

In the following sections we simplify the notation by leaving out conditioning on the known parameters \( \mathbf{x}_t \) and \( \mathbf{e}_t \).

1) Sequential Bayesian Filtering: Let us assume the joint filtering distribution \( p(\mathbf{a}_{t-1}, \theta | \mathbf{y}_{1:t-1}^{MA,SID}) \) at time \( t - 1 \) is known from the previous update step. Then, we compute the predictive distribution as:

\[
p(\mathbf{a}_t, \theta | \mathbf{y}_{1:t}^{MA,SID}) = \sum_{\mathbf{a}_{t-1}} p(\mathbf{a}_t | \mathbf{a}_{t-1}) p(\mathbf{a}_{t-1}, \theta | \mathbf{y}_{1:t-1}^{MA,SID}).
\]

Since there is no observation history at the time \( t = 1 \), we initialize the update process by the following distribution:

\[
p(\mathbf{a}_1, \theta) = \begin{cases} \prod_{k=1}^{L_1} p(a_k, \theta) \frac{1}{|\Theta|} & \text{all allowed } \theta, \\ 0 & \text{otherwise}. \end{cases}
\]

The number of possible trajectory-to-identity associations is \(|\Theta| = I(I-1)\ldots(I-L+1)| and the size of the joint state-parameter space is \(2^{K_t} |\Theta|\).

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:

\[
p(\mathbf{a}_t | \mathbf{y}_{1:t}^{MA,SID}) \sim \sum_{\theta} p(\mathbf{y}_t^{MA,SID} | \mathbf{a}_t, \theta) p(\mathbf{a}_t, \theta | \mathbf{y}_{1:t-1}^{MA,SID}).
\]

After the new observation \( \mathbf{y}_t^{MA,SID} \) is introduced we can obtain the updated distribution of the unknown parameter \( p(\theta | \mathbf{y}_t^{MA,SID}) \) by marginalization over the state space:

\[
p(\theta | \mathbf{y}_1:t^{MA,SID}) \sim \sum_{\mathbf{a}_t} p(\mathbf{y}_t^{MA,SID} | \mathbf{a}_t, \theta) p(\mathbf{a}_t, \theta | \mathbf{y}_{1:t-1}^{MA,SID}).
\]

Finally, since on the interval \([1, t]\) only \(L_{1:t}\) trajectories were registered, the meaningful part of the parameter distribution is \(p(\theta_{1:L_{1:t}} | \mathbf{y}_1:t^{MA,SID})\) and we obtain it by the marginalization over \(\theta_{L_{1:t}+1:L}\).

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

for \( t = 1 \) to \( T \) do
(A) Compute \(p(\mathbf{a}_t, \theta | \mathbf{y}_{1:t-1}^{MA,SID})\):

if \( t = 1 \) then
  Use Equation (6).
else
  Use Equation (5).
end if
(B) Compute \(p(\mathbf{a}_t | \mathbf{y}_{1:t}^{MA,SID})\). Use Equation (7).
(C1) Compute \(p(\theta | \mathbf{y}_{1:t}^{MA,SID})\). Use Equation (8).
(C2) Compute \(p(\theta | \mathbf{y}_{1:t}^{MA,SID})\):

\[
p(\theta_{1:L_{1:t}} | \mathbf{y}_1:t^{MA,SID}) = \sum_{\theta \in \Theta_{L_{1:t}+1:L}} p(\theta | \mathbf{y}_{1:t}^{MA,SID})
\]

end for

The total complexity of the forward sequential filtering algorithm on interval \([1, T]\) is \(O(T \cdot 2^{2L} \cdot |\Theta|)\).

2) Viterbi Decoding: In this algorithm we use the same update Equations (6), (5) and (8) to get the distribution of the unknown association parameters. The main difference is that at the final time \(T\) the total number of the detected trajectories and information about their overlap in time is available. Therefore, at \(T\) the set of all possible trajectory-to-identity associations \(\Theta\) is known exactly. We incorporate this information in the Equation (6) which allows us to work with valid associations only and decreases the overall computational complexity.

**Algorithm 2** Viterbi Decoding

for \( t = 1 \) to \( T \) do
(A) Compute \(p(\mathbf{a}_t, \theta | \mathbf{y}_{1:t-1}^{MA,SID})\):

if \( t = 1 \) then
  Use Equation (6).
else
  Use Equation (5).
end if
(B) Compute \(p(\theta | \mathbf{y}_{1:t}^{MA,SID})\). Use Equation (8).
(C) Find optimal parameter \(\theta^*\). Use Equation (4).
(D) Find optimal state sequence \(\mathbf{a}_{1:T}^*\). Use Equation (4).

end for

Steps (C) and (D) have complexity \(O(T \cdot 2^L)\) and therefore, the 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 it being overlapped speech, where the average durations of the segments with one and two active speakers are respectively 8.4s and 3.3s. Since both datasets contain read speech with comparable speakers’ energies no difference in performance was noticed and we present joint results for both datasets.

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 omni-directional 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. Since the extraction of the local maxima is a somewhat ill defined problem we choose to filter extracted regional maxima and to present dependence of the segmentation performance on different threshold values. Threshold values 55%, 75% and 100% were chosen since we want to present performance over wider interval of threshold values, and on intervals [55%, 75%] and [75%, 100%] we noticed, respectively, performance increase and decrease. 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. The first set contains all extracted regional maxima; the second set, all regional maxima grater than 55% of the global maximum; and the 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.

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. This gives insight into the dependency of the false speaker detections on the 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 $p_{fd}$ rise with relaxation of the neighborhood size $A$ while the expected number of false alarms $\lambda$ 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 remaining experiments for the threshold levels 55% and 75%.

<table>
<thead>
<tr>
<th>$A$</th>
<th>$R_{LM} &gt; 0$</th>
<th>$R_{LM} &gt; 0.55R_{GM}$</th>
<th>$R_{LM} &gt; 0.75R_{GM}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>8.5</td>
<td>34.8</td>
<td>54.8</td>
</tr>
<tr>
<td>10</td>
<td>21.7</td>
<td>64.6</td>
<td>83.9</td>
</tr>
<tr>
<td>15</td>
<td>39.0</td>
<td>86.5</td>
<td>97.5</td>
</tr>
<tr>
<td>20</td>
<td>59.4</td>
<td>93.2</td>
<td>97.7</td>
</tr>
<tr>
<td>25</td>
<td>76.3</td>
<td>95.5</td>
<td>97.6</td>
</tr>
<tr>
<td>30</td>
<td>86.6</td>
<td>97.8</td>
<td>97.6</td>
</tr>
</tbody>
</table>

Observations for the speaker identification are MFCC coefficients extracted on 100ms frames aligned with microphone array frames. Having in mind results presented in [19], properties of our dataset (6 speakers, 30sec training segments, 8.5/3.5/sec average single/overlapped speaker turn duration) and results from preliminary speaker recognition experiments, we decided to train 16-component GMMS with diagonal covariance matrices. Models for silence and single participants were trained on 30s training samples, while GMMS for two overlapped participants were trained on 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 standard for speaker segmentation type of problems [4], [31] and gives more insight into the performance than using just the number of correctly detected speaker-frames.

\[
P = \frac{\# \text{ of found true active speaker-frames}}{\# \text{ of found active speaker-frames}} (9)
\]

\[
R = \frac{\# \text{ of found true active speaker-frames}}{\# \text{ of true active speaker-frames}} (10)
\]

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

- $\text{MA}_{0.55}$ and $\text{MA}_{0.75}$ denote microphone array likelihood models, $\alpha, \beta = (1, 0)$, with observations obtained with threshold levels 55% and 75% respectively.
- $\text{MA}_{0.00}$ denotes the baseline microphone array likelihood 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 the $F$-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 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, $p(a_{t+1} = a_i|a_t) = 0.99$, 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 $p(a_{t+1} = a_i|a_t) = 0.99$ which we used throughout the experiments we present Tables II and III. All values in these tables are obtained for the best likelihood model $MA_{0.75}$&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 II. PERFORMANCE VS. STATE TRANSITION MODEL PARAMETER $p(a_{t+1} = a_i|a_t)$ FOR NEIGHBORHOOD SIZE $A = 10$ |
|------------------|------------------|------------------|
| $p_{i, t}$      | $P_C$            | $P_T$            |
| 0.1             | 0.8              | 0.3              |
| $P_T$           | 94.6             | 93.1             |
| $R_T$           | 97.6             | 96.1             |
| $F_T$           | 92.9             | 91.3             |

| TABLE III. PERFORMANCE VS. NEIGHBORHOOD SIZE $A$ FOR TRANSITION PARAMETER $(p_{i, t} = 0.99)$ |
|------------------|------------------|------------------|
| $A_{i, t}$       | 0.10             | 0.15             |
| $P_T$           | 94.6             | 91.4             |
| $R_T$           | 98.7             | 94.4             |
| $F_T$           | 96.3             | 93.6             |

For both decoding schemes, our likelihood models ($MA_{0.55}$ and $MA_{0.75}$) give better overall $F$-measure performance than the baseline ($MA_{1.00}$). When employing the Viterbi decoding scheme (Table IV) on segments with only one active speaker, the 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%). The relative improvement over the baseline on segments with overlapped speech is 46.7%.

Tables VI and VII 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.

| TABLE IV. SEGMENTATION PERFORMANCE: VITERBI DECODING, NO LIMIT ON NUMBER OF ACTIVE SPEAKERS |
|------------------|------------------|------------------|------------------|------------------|------------------|
| $P_T$            | 0.75             | 0.75             | 0.75             | 0.75             | 0.75             |
| $P_T$            | 96.7             | 96.4             | 95.8             | 95.5             | 95.2             |
| $R_T$           | 99.3             | 99.6             | 99.4             | 99.3             | 99.1             |
| $F_T$           | 98.5             | 98.8             | 98.6             | 98.9             | 98.5             |

| TABLE V. SEGMENTATION PERFORMANCE: BAYESIAN FILTERING, NO LIMIT ON NUMBER OF ACTIVE SPEAKERS |
|------------------|------------------|------------------|------------------|------------------|------------------|
| $P_T$            | 0.75             | 0.75             | 0.75             | 0.75             | 0.75             |
| $P_T$            | 98.2             | 98.9             | 99.7             | 99.4             | 99.1             |
| $R_T$           | 99.2             | 98.8             | 99.1             | 99.6             | 99.3             |
| $F_T$           | 98.4             | 99.5             | 99.3             | 99.3             | 99.1             |

We tested the 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 Tables IV and V. 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. The last three columns contain performance measures on segments with overlapped speech.
Similarly, for the transition model that poses no limit on the number of active speakers, the proposed MA likelihood models perform better than the baseline ones.

The proposed modality fusion model brings further performance improvement. This validates our assumption on the complementarity of SID and MA likelihood models. Note that the combination of the baseline MA\(_{1.00}\) and the SID likelihoods degrades MA\(_{1.00}\) performance on the segments with single active speaker and improves performance on the segments with overlapped speech. On the other hand, combination of SID with our models MA\(_{0.55}\) and MA\(_{0.75}\) 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 [19] 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. The Viterbi decoding of the optimal sequence provides a better performance and should be the method of choice for applications where batch segmentation is possible. The presented results show that in situations where sequential and real time processing is necessary forward Bayesian decoding provides slightly lower performance while preserving trends presented and discussed for the Viterbi decoding scheme.

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 \(y_t = (y_{t,1}^{MA}, \ldots, y_{t,K_t}^{MA})\), locations of the participants \(x_t = (x_{t,1}^{1}, \ldots, x_{t,K_t}^{T})\) and the speaker activity vector \(a_t = (a_{t,1}, \ldots, a_{t,K_t})\) the joint likelihood is defined as:

\[
p(y_t^{MA}|x_t, a_t) = \prod_{i:a_{t,i}=1} p_1(y_t^{MA}|x_{t,i}) \prod_{i:a_{t,i}=0} p_0(y_t^{MA}|x_{t,i}),
\]

where the probabilities \(p_1\) and \(p_0\) are given by the following equations:

\[
p_1(y_t^{MA}|x_{t,i}) = \begin{cases} L_1(\langle j \rangle : ||y_t^{MA} - x_{t,i}|| \leq A) & \text{if } \exists j : ||y_t^{MA} - x_{t,i}|| \leq A, \\ L_0(\forall j : ||y_t^{MA} - x_{t,i}|| > A) & \text{otherwise}, \end{cases}
\]

\[
p_0(y_t^{MA}|x_{t,i}) = \begin{cases} L_1(\langle j \rangle : ||y_t^{MA} - x_{t,i}|| > A) & \text{if } \exists j : ||y_t^{MA} - x_{t,i}|| > A, \\ L_0(\forall j : ||y_t^{MA} - x_{t,i}|| \leq A) & \text{otherwise}. \end{cases}
\]

The constant \(A\) defines the 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 the ratio \(\frac{L_1}{L_0}\) has to be significantly greater than one.

ACKNOWLEDGMENT

This research was supported in part by grants from the National Science Foundation, ONR and the Army.

REFERENCES


Viktor Rozgic received Dipl.Eng. degree in electrical engineering from the University of Belgrade, Serbia, in 2001, and the M.S. degree from the University of Southern California, Los Angeles, in 2007. He is currently a Ph.D. candidate in the Signal Analysis and Interpretation Laboratory at the University of Southern California, Los Angeles. The focus of his Ph.D. dissertation is on the multi-modal monitoring and analysis of multi-party interactions.

His general research interests include audio-visual signal processing, multi-modal fusion, multimedia content analysis, sequential and Markov chain Monte Carlo filtering methods, and multi-target tracking algorithms.

Kyu J. Han (S’00) received the B.S. degree in radio communications engineering from Yonsei University, Seoul, Korea, in 2000, and the M.S. degree in electrical engineering from Seoul National University, Seoul, in 2002. He is currently pursuing the Ph.D. degree in electrical engineering at the University of Southern California, Los Angeles. His dissertation focus is on robust speaker clustering under data source variation for reliable speaker diarization.

His general research interests include speech and speaker recognition, audio signal processing, and multimedia content analysis.

Panayiotis G. Georgiou (M’02) received the B.A. and M.Eng. degrees from Cambridge University (Perbroke College), U.K., in 1996 and the M.Sc. and Ph.D. degrees from the University of Southern California (USC), Los Angeles, in 1998 and 2002, respectively. From 1992 to 1996, he was awarded a Commonwealth scholarship from Cambridge-Commonwealth Trust, Since 2003, he has been a member of the Speech Analysis and Interpretation Lab, USC, first as a Research Associate and currently as a Research Assistant Professor. His interests span the fields of human social and cognitive signal processing. He has worked on and published over 40 papers in the fields of statistical signal processing, alpha stable distributions, speech and multimodal signal processing and interfaces, speech translation, language modeling, immersive sound processing, sound source localization, and speaker identification.

His current focus is on multimodal cognitive environments and speech-to-speech translation.

Shriram Narayanan is Andrew J. Viterbi Professor of Engineering at the University of Southern California (USC), and holds appointments as Professor of Electrical Engineering and jointly as Professor in Computer Science, Linguistics and Psychology. He received his Ph.D. in Electrical Engineering from UCLA in 1995. Prior to USC he was with AT&T Bell Labs and AT&T Research, first as a Senior Member, and later as a Principal member, of its Technical Staff from 1995-2000. At USC he is a member of the Signal and Image Processing Institute and directs the Signal Analysis and Interpretation Laboratory. His research focuses on human-centered information processing and communication technologies. Shri Narayanan is an Editor for the Computer Speech and Language Journal (2007-present) and an Associate Editor for the IEEE Transactions on Multimedia and for the Journal of the Acoustical Society of America (2009-present). He was also an Associate Editor of the IEEE Transactions of Speech and Audio Processing (2000-04) and the IEEE Signal Processing Magazine (2005-2008). He served on the Speech Processing technical committee (2005-2008) and Multimedia Signal Processing technical committees (2004-2008) of the IEEE Signal Processing Society and presently serves on the Speech Communication committee of the Acoustical Society of America and the Advisory Council of the International Speech Communication Association. Shri Narayanan is a Fellow of the Acoustical Society of America, a Fellow of IEEE, and a member of Tau-Beta-Pi, Phi Kappa Phi and Eta-Kappa-Nu. He is a recipient of an NSF CAREER award, USC Engineering Junior Research Award, USC Electrical Engineering Northrop Grumman Research Award, a Provost fellowship from the USC Center for Interdisciplinary research, a Mellon Award for Excellence in Mentoring, an IBM Faculty award, an Okawa Research award, and a 2005 Best Paper award from the IEEE Signal Processing society (with Alex Potamianos). Papers with his students have won best paper awards at ICSLP’02, ICASSP’05, MMSP’06, MMSP’07 and DCOSS’09. He has published over 350 papers and has seven granted U.S. patents.