Exploring Co-occurence between Speech and Body Movement for Audio-guided Video Localization

Himanshu Vajaria, *Member, IEEE*, Sudeep Sarkar, *Senior Member, IEEE*, and Rangachar Kasturi, *Fellow, IEEE*

***Abstract*—This paper presents a bottom-up approach that combines audio and video to simultaneously locate individual speakers in the video (2-D source localization) and segment their speech (speaker diarization), in meetings recorded by a single stationary camera and a single microphone. The novelty lies in using motion information from the entire body rather than just the face to perform these tasks, which permits processing non-frontal views unlike previous work. Since body-movements do not exhibit instantaneous signal-level synchrony with speech, the approach targets long term co-occurrences between audio and video subspaces. First, temporal clustering of the audio produces a large number of intermediate clusters, each containing speech from only a single speaker. Then, spatial clustering is performed in the video frames of each cluster by a novel eigen-analysis method to find the region of dominant motion. This region is associated with the speech assuming that a speaker exhibits more movement than the listeners. Thus partial diarization and localization is obtained from the intermediate clusters. Speech from an intermediate cluster is modeled by a mixture of Gaussians and the speaker's location is represented by an eigen-blob model. In the ensuing iterative clustering stage, the diarization and localization results are progressively refined by merging the closest pair of clusters and updating the models until a stop criterion is met. Ideally, each final cluster contains all the speech from a single speaker and the corresponding eigen-blob model localizes the speaker in the image. Experiments conducted on 21 hours of real data indicate that the proposed localization approach leads to a relative improvement of 40% over Mutual Information based localization and that speaker diarization improves by 16% by incorporating visual information. The proposed approach does not require training and does not rely on *a priori* hand/face/person detection.**

***Index Terms*—Audio-visual association, Meeting analysis, Speaker localization, Speaker diarization.**

I. Introduction

MEETINGS are an integral part of our daily lives, where information is disseminated, ideas are discussed and decisions are taken. Consequently, many organizations have begun archiving their meetings for future review. However, to be of practical use, these large and constantly growing archives should be comprehensively indexed so that they may support a variety of queries such as query for a discussion topic, or for an individual's comments, or for specific activities such as presentations and note-taking. Determining who spoke when (speaker diarization) and locating the current speaker (speaker localization) are prerequisites for such queries, as well as for higher level tasks such as generating audio-visual summaries and meeting transcripts.

The semantic analysis of meetings is receiving considerable interest, sparking evaluations such as the ones by NIST [7] and CLEAR [22], where meetings are recorded in special rooms rigged with multiple microphones and cameras. However, this work focuses on meetings recorded by a simple setup consisting of a *single camera* and a *single microphone*, because of its broader applicability. As everyday devices such as laptops, PDAs and cell phones have become capable of video recordings, such devices can be used to record a group meeting, effectively converting any location into a meeting room. Also, techniques developed for this constrained setup can be used for surveillance applications, where covertness requires using a simple portable recorder.

In previous work on speaker diarization and localization in the single camera, single microphone scenario, the problem is posed as one of detecting synchronous audio-visual events. Mutual information (MI) based approaches have been successfully demonstrated in situations where the faces are frontal and have a high resolution. Since the speaker's face and lips are clearly visible when speech is heard, an instantaneous synchrony exists between the audio and video, which is successfully exploited by MI based approaches.

However, meeting room videos are quite different as multiple persons are seated facing each other and not the camera. Thus the faces are not necessarily frontal. Also, since the camera is placed much farther from the participants, faces have a low resolution. As a result, a person's lips may not be clearly visible when they speak. Additionally, participants often exhibit a high degree of movement for short intervals even when they do not speak such as when taking notes, sipping coffee, or swiveling in a chair, and such movements are falsely associated with the speech. For these reasons, we find that MI based approaches do not perform well on meeting datasets [24].

We propose a different framework for audio-visual integration motivated by the following observations. A strong synchrony exists between the lip movements of a speaker and the resultant speech which has been exploited in MI based works. There also exists a loose association between a person's speech and head/hand gestures which has been demonstrated in works such as [19], [25]. In addition to the relation of speech with lips and gestures, we observe that in general a person exhibits more movement during speech. To maintain eye contact, the head turns from one listener to the other and usually bobs up and down during speech because of jaw movements. Also, the speaker's hands and shoulders move

The work was supported in part by the USF Computational Tools for Discovery Thrust.

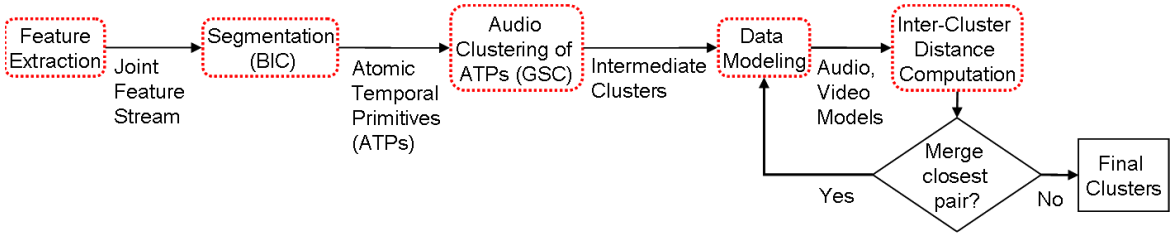

Fig. 1. System Flowchart: Audio and video features are fused to form a joint feature stream which is partitioned using the Bayesian Information Criterion (BIC) into atomic temporal primitives (ATPs). ATPs are grouped into intermediate clusters by Graph Spectral Clustering. This is followed by an iterative clustering stage that produces one cluster per speaker.

involuntarily when an idea is expressed.

Such movements are not synchronized with speech, i.e. there does not exist an instantaneous mapping between audio and video features. Rather, there exists a long term co-occurrence of speech and movement, i.e. over longer durations, people exhibit more body movements when speaking than when listening. We exploit this phenomenon of co-occurrence of speech and body movements to perform speaker diarization and localization, assuming that in general a speaker moves more than a listener.

The flowchart in Figure 1 illustrates our approach. Audio and video features are concatenated to obtain a joint feature stream. The Bayesian information criterion (BIC) finds changepoints in this stream, which are frames where there is a discontinuity in the audio-visual pattern, signaling a change in speaker. These changepoints partition the stream into contiguous atomic temporal primitives (ATPs) that are of short durations and have homogeneous audio-visual characteristics.

In the next step, graph spectral clustering (GSC), groups together ATPs based on their audio content into intermediate clusters, which are further processed in an iterative framework. Audio and video models are built from the intermediate clusters and used to compute distances between each pair of clusters. In each iteration, the closest pair of clusters are merged and new models are built from it. Since the merged cluster is of longer duration, a video model built from it will lead to better speaker localization, which in turn will positively influence the clustering procedure in the next iteration. This clustering-modeling cycle continues till a stop criterion is met, resulting in the final clusters. Ideally, each of the final clusters contains all the speech from a single person, effectively performing speaker diarization and the cluster's video model localizes the speaker in the video.

The outline of the rest of this paper is as follows. The creation of a joint feature stream and its partitioning into ATPs is described in section III. Section IV deals with clustering ATPs into intermediate clusters and the iterative audio-visual clustering framework. Section V presents the improvements in diarization by incorporating video and compares eigenblob and MI based localization results. Section VI carries the conclusions while the next section surveys related work.

II. Related Work

Person detection/tracking and speaker diarization are tasks that have been heavily studied by the computer vision, and signal processing communities, respectively. Lately, there has been much emphasis on integrating audio and video to jointly perform these tasks in meeting rooms [3], [6], [11]. Programs such as the *Rich Transcription evaluation (RT)* [7] and the *CLassification of Events Activities and Relationships (CLEAR)* [22] aim to further such research. Their focus is on data collected in *smart rooms*, rigged with multiple sensors. Additionally, there is significant work on diarization and localization in the single camera and/or single microphone scenario, which we review here.

Speaker diarization and localization has been performed using only a single camera when the speaker's face is frontal, unoccluded, and exhibits more movement than other faces in the image. The approach involves using a face detector to locate all faces in the image and then using motion in the region around the face [14] or mouth [17] to determine the speaker. Speaker diarization using a single audio channel relies on the phenomenon that speech from different persons have different spectral characteristics and diarization is performed by unsupervised clustering of audio features [23].

The multimodal approach (single camera and single microphone) has usually been demonstrated in scenarios where the speakers are facing the camera such as in broadcast news videos or on the CUAVE [16] database. As a result, the faces are frontal, and an instantaneous synchrony between the audio and video signals exists. The problem is typically formulated as finding projections that maximize the mutual information between the projected audio and video signals. Works in this category [5], [8], [10], [12], [21] differ based on the choice of audio and video representations, whether or not the features are projected onto a learned subspace before modeling, and the paradigm used to model the audio, video, and joint signals.

The audio signal is typically represented by MFCCs, LPCs, or a spectrogram. The video signal is represented by image intensity, image differences, or DCT coefficients. The audio and video features are either modeled directly or training data is used to learn an optimal subspace that maximizes synchrony between the projected features. Either parametric (usually Gaussian) or non-parametric models have been used to model the signals. In works such as [1], [13], audio-visual association is performed based on short-term co-occurrences between audio and video primitives.

All of the above works model relationships between the audio and video signals by finding image regions that are synchronized with the audio. Their assumption is that the

underlying cause which produces the audio and video signals, always expresses itself in both modalities and that this relationship is instantaneous. Although the assumption holds for cases where two people are facing the camera and taking turns at speaking or when the object generating the sound is visible, it does not hold for meetings captured by a single camera as all faces are not frontal. This leads to the poor performance of such approaches for meeting scenarios.

Audio-only speaker diarization involves unsupervised clustering of audio features and video-only person detection involves clustering pixels in the images space based on appearance and/or motion models. On the other hand, audio-visual synchrony methods seek correlations between speech features and image pixels, without explicitly clustering in individual sub-spaces. The proposed approach seeks to combine these three facets - by first over-clustering in the audio subspace to find longer temporal durations when a person is speaking. Next, clustering is performed in the video-space by grouping pixels with high covariance in frames from these durations. The audio and video clusters thus obtained are associated with the same person, assuming that the dominant motion in video frames is due to the speaker's movements.

III. Atomic Temporal Primitives

The first step of our approach involves partitioning the meeting into contiguous durations that we term as Atomic Temporal Primitives (ATPs). The ATPs should be homogeneous - i.e. each ATP should contain speech and movement from only a single speaker. This partitioning step is very similar to the segmentation task in speaker diarization, where changepoints are sought in the audio stream that indicate a change in speaker. The BIC framework [4] which has proven effective for audio segmentation, is used in this work to find ATP boundaries in the joint-feature stream.

However, we also incorporate video information to detect such changepoints, motivated by the following observation: A change in speaker is indicated by a change in the model producing the audio features which is the premise of the audio-based BIC approach. Often times, a change in speaker is also reflected by a change in the video dynamics. After a person stops speaking, they often change posture - by leaning back further into their chair indicating through a non-verbal mechanism that the floor is open. Similarly, just prior to speaking, a person attempts to gain their audience's attention by leaning forward or extending their arm into the common space to indicate a desire to hold the floor. Thus a change in speaker is also reflected by a change in the image regions where motion occurs and this phenomenon can be exploited to detect speaker changepoints.

Prior to performing segmentation, a speech/silence detector is run to eliminate durations of silence from the recording. The elimination of silence frames is necessary as video information during silence adversely affects the segmentation performance since motion during these frames is spurious in nature and not related to speech activity . Secondly, since some meetings may have extended durations of silence, eliminating these frames, speeds up processing. After eliminating silence segments, ATP boundaries are found from a joint feature stream produced by concatenating audio and video features. Mel-Frequency Cepstral coefficents (MFCCs), are used as the audio features. The MFCCs are extracted using 32 filters with the bandwidth ranging from 166 Hz to 4000 Hz. The MFCCs ($\mathcal{A}$) are then projected onto a PCA space to obtain a low dimensional representation (A).

The video features which intend to capture motion, are obtained using image differences (three frames apart). The difference images are thresholded to suppress jitter and dilated by a 3×3 circular mask to enhance regions of motion. The images are then downsampled from their original size of 480×720 to 48×72 and vectorized. The video features ($\mathcal{V}$) are then projected onto their PCA space to obtain their projections (V). A joint audio-visual subspace is obtained by concatenating the projections using

$$X(t) = \begin{bmatrix} s_f \cdot A(t) \\ V(t) \end{bmatrix} \quad (1)$$

Here $A(t) = [A_1(t), A_2(t), \ldots, A_{d_A}(t)]^T$, where $A_1(t), A_2t, \ldots, A_{d_A}(t)$ are the PCA coefficients of the audio features. Similarly, $V(t) = [V_1(t), V_2(t), \ldots, V_{d_V}(t)]^T$, where $V_1(t), V_2t, \ldots, V_{d_V}(t)$ are the PCA coefficients of the video features. The index t represents the frame number, d_A and d_V represent the dimensionality of the audio and video features, respectively and $d_X = d_A + d_V$ is the dimensionality of the resulting joint feature (X). In our experiments, d_A and d_V were chosen as 8 and 24, respectively, retaining 90% of the original variance. The scaling factor s_f is set to $\sqrt{|\Sigma_V|/|\Sigma_A|}$, where Σ_A and Σ_V are the covariances of the audio and video features, respectively. The scaling ensures that both features contribute equally to the joint feature stream.

The joint feature stream is the partitioned into ATPs using the Bayesian Information Criterion (BIC). For the mathematical and implementation details of BIC, we refer the reader to [24], and provide an intuitive explanation here. The BIC based segmentation operates on the principle that a sudden change in the feature space is caused by a change in the underlying model. A change in speaker, implies a change in the audio model. Also, as mentioned earlier, there will be a change in the image region where motion occurs. Since the difference images are projected as a low-dimensional vector and modeled by a unimodal multivariate Gaussian (across time), a change in the image region will be modeled by a different Gaussian model. The joint Gaussian is more sensitive to speaker changes than models built for either audio or video alone. This however comes at the cost of increased false detections due to the video - such as when a person reaches out to grab a cup when someone else is speaking. However, since ATPs can be merged in the clustering stage, false detects are not as expensive as missed detects.

IV. Clustering and Localization

Once the feature stream has been split into ATPs, the next goal is to merge all ATPs containing speech from the same individual. The localization task involves determining the image region in the video frames of those ATPs where

the speaker is seated. These two tasks can be performed sequentially - speaker diarization can be performed first using only the audio and then video frames from the final clusters can be analyzed to locate the speaker. Alternatively, since the video contains information about the current speaker, both audio and video features can be used from the ATPs to jointly perform diarization and localization.

However, since individual ATPs tend to be of short durations, the visual information in them is not very consistent. For example, where a person utters just a few sentences, we observe that there is little accompanying motion and that this situation exacerbates when the person is facing away from the camera. Similarly an ATP can contain speech from one person but motion from more than one individual - as occurs when someone is taking notes. The hypothesis on which this work is based, is that on an average, a speaker exhibits more movement than a listener and this holds when considering longer time durations. Thus, instead of obtaining video models from the ATPs, the ATPs are first clustered using only the audio to obtain fewer large clusters. Video models can be reliably estimated from these larger intermediate clusters, and then be used to influence diarization in the iterative diarization-localization process.

The rest of this section is structured as follows. Subsection IV-A, describes the grouping of ATPs into intermediate clusters. Subsection IV-B, deals with modeling the audio and video features of these intermediate clusters and subsection IV-C describes the iterative diarization-localization procedure.

A. Intermediate Clusters

The initial clustering of ATPs is performed using only audio by modeling its MFCCs by a unimodal Gaussian with a full covariance matrix. The clustering problem is formulated as a graph partitioning problem. Each ATP is represented as a node and the ΔBIC distance distance [26] between each pair of ATPs, serves as the edge weights to obtain a completely connected graph. A recursive graph bi-partitioning algorithm [20] is then used to group ATPs into sixteen clusters, motivated by the observation that meetings usually contain less than sixteen speakers and so the data is not under-clustered. Also at sixteen clusters, we find for our dataset that each intermediate cluster contains sufficient data to robustly estimate audio and video models.

B. Audio and Video Models

Once the ATPs have been grouped into intermediate clusters, audio models are built from them using the UBM-GMM technique described in [18]. The features used for building the models are the PCA projections of the MFCCs (A). In this technique, first a Universal Background Model (UBM) is built using the entire speech in the meeting. The UBM is essentially a Gaussian mixture model with K mixtures, $\Phi^u = \{\omega_k^u, \mu_k^u, \Sigma_k^u\}$, where ω_k^u represent the weights (with the constraint $\sum_{k=1}^{K} \omega_k^u = 1$), μ_k^u represents the d_A dimensional mean vectors and Σ_k^u are the $d_A \times d_A$ covariance matrices. In our implementation, the UBM consists of eight Gaussians with diagonal covariance matrices learned using the Expectation Maximization (EM) algorithm.

From this UBM, we obtain GMMs $\Phi^i = \{\omega_k^i, \mu_k^i, \Sigma_k^i\}$, for each intermediate audio cluster a_i, by adapting only the means of Φ^u by a *maximum a posteriori* (MAP) adaptation [2]. Since means of the k^{th} component of all intermediate clusters are adapted from the same mean (μ_k^u), there exists a one-one correspondence between them. This allows us to efficiently compute the distance between the audio models of two intermediate clusters, a_i and a_j as

$$d(a_i, a_j) = \sqrt{\sum_{k=1}^{K} \omega_k^u (\mu_k^i - \mu_k^j)^T (\Sigma_k^u)^{-1} (\mu_k^i - \mu_k^j)} \qquad (2)$$

where μ_k^i and μ_k^j represent the means of the k^{th} component of intermediate audio clusters a_i and a_j, respectively, and T is the transpose operator. It was shown in [2] that this distance is highly correlated with a Monte Carlo estimation of the KL2 distance, with the added advantage that it is much cheaper to compute.

The intermediate clusters also serve as the starting point for speaker localization. A video model is built from the video features ($\mathcal{V}$) of each intermediate cluster by analyzing the eigenvectors of its video features. Let v_i represent the set of video features from an intermediate video cluster and let Σ_{v_i} represent its covariance. Solving

$$\Sigma_{v_i} E = \Lambda E \qquad (3)$$

we obtain the eigen-vectors of Σ_{v_i} as the column entries of E, where Λ is the corresponding eigen-value matrix.

Since eigen-vectors are projections that reduce the covariance of the projected variables, they effectively group pixels that move together. If the dominant speaker moves the most in the set of frames, the primary eigenvector partitions the image into two regions - one belonging to the speaker and the other to spurious background movements. However, it cannot be determined which of the two regions corresponds to the speaker from only the primary eigenvector. Since the second eigenvector is orthogonal to the first, it splits the dominant component of the first eigenvector - which is the region that represents the speaker's location.

Mathematically, if e_1 is the largest eigenvector and e_2 is the second largest eigenvector of frames from the intermediate video cluster v_i, then the part r_i which represents the selected region of e_1 is given by

$$r_i = \left\{ \begin{array}{ll} |e_1^+| & \text{if} \quad |e_2^T e_1^+| < |e_2^T e_1^-| \\ |e_1^-| & \text{if} \quad |e_2^T e_1^-| < |e_2^T e_1^+| \end{array} \right\} \qquad (4)$$

where e_1^+ and e_1^- are the positive and negative parts of the primary eigenvector, and T is the transpose operator. The dominant region r_i is then normalized so that it sums to unity and serves as the eigen-blob model for the video cluster v_i. This eigen-blob model (r_i) is basically a probability density function representing the likelihood of a pixel belonging to the speakers location.

Figure 2 illustrates the eigen-blob localization for two intermediate clusters from a meeting of four people. The

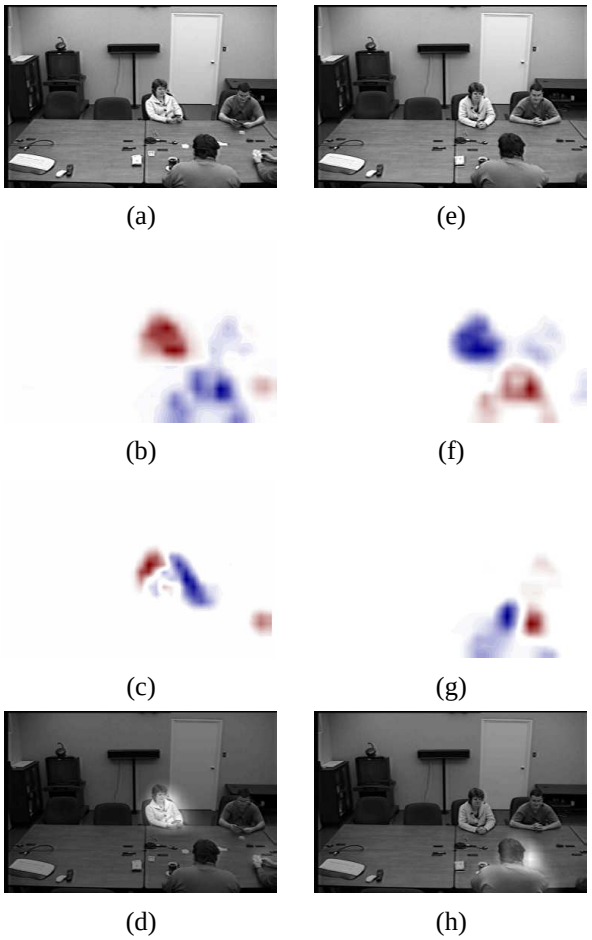

Fig. 2. Localization by the eigen-blob method. (a) and (e) show sample images from two intermediate clusters in which the speaker is located at the top-left and bottom-left, respectively. (b) and (f) show the respective principal eigenvectors. The second eigenvectors shown in (c) and (g), split the dominant region (positive or negative) of the primary eigenvectors. The dominant regions shown in (d) and (h) represent the speaker's location.

first eigenvector has non-zero components corresponding to the moving parts of the image; in addition, the sign of the eigenvector further divides the moving portions into two parts (shown by two different colors). The second eigenvector, which captures the next dominant mode of motion correlation and is orthogonal to the first eigenvector, is used to identify the portion from the speaker.

Intermediate clusters belonging to the same speaker should have similar video characteristics. Specifically, the eigen-blob models should overlap, and the degree of overlap can be considered as a measure of similarity. Since the eigen-blob models are non-parametric densities signifying the speaker's location within the image, the distance between two models is computed using the symmetric Kullback-Leibler (KL2) measure as

$$d(v_i, v_j) = \frac{1}{2}\left(\sum_{\alpha} r_i(\alpha)\log\frac{r_i(\alpha)}{r_j(\alpha)} + \sum_{\alpha} r_j(\alpha)\log\frac{r_j(\alpha)}{r_i(\alpha)}\right) \quad (5)$$

where α is the variable that spans the eigen-space.

As a comparison to the eigen-blob localization approach, we also implemented the mutual information (MI) based localization technique. The MI between two multivariate random

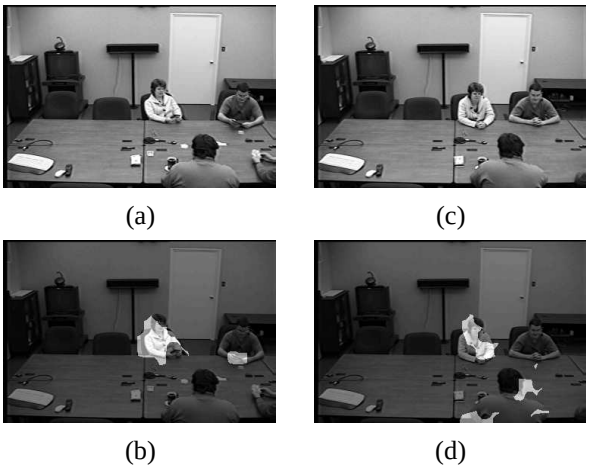

Fig. 3. Localization using mutual information (MI). (a) shows a sample image where the person in the top left is speaking and (c) shows a sample image where the person on the bottom left is speaking. (b) and (d) show the MI images for (a) and (c), respectively.

variables $\mathcal{X}$ and $\mathcal{Y}$ is given by

$$I(\mathcal{X};\mathcal{Y}) = \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} \log\left(\frac{p(x,y)}{p(x)p(y)}\right) \quad (6)$$

where $p(x,y)$ is the joint probability distribution and $p(x)$ and $p(y)$ are the marginal distributions.

Similar to [14], we compute the MI between the audio features A, and each pixel $\mathcal{V}_{x,y}$ of the video feature $\mathcal{V}$. The MI is computed every frame using a two second window to estimate the probability distributions $p(x)$, $p(y)$ and $p(x,y)$ which are assumed to be Gaussian.

Figure 3 illustrates sample results of localization using the MI approach. The MI is computed between the audio and each image pixel. Pixels that are highly synchronized with the audio have a higher MI value. The MI image is thresholded to discard low value pixels and the filtered image is displayed in (b) and (d). The localization output is considered as the connected component with the largest average MI.

The representative MI localization images in Figure 3 show that MI performs better when the speaker is facing the camera as seen in (b) than when the person is facing away from the camera as in (d). Interestingly, (d) shows that even when the face is not visible, there are regions around the speaker's body that are associated with the speech. Compared to Figure 3 (b) and (d), the localization results are better in Figure 2 (d) and (h). We believe that this is because the MI approach seeks instantaneous associations between the pixels and the speech - a relation which is non-robust in the meeting domain whereas the eigen-blob approach seeks correlated pixels in frames that are determined to belong to the same speaker using the audio channel. Since the eigen-blob approach considers longer durations, spurious movements by non-speakers are averaged out leading to better localization results.

C. Iterative Clustering

Once audio and video models have been built, the intermediate clusters are merged using an agglomerative clustering

framework in which the audio and video models are refined at each iteration. The procedure involves merging the closest pair of clusters into a new cluster and obtaining a new audio and a new video model for that cluster. Since the merged cluster contains more data than either of the individual clusters, models derived from it would be more robust and representative of the speaker's audio-visual characteristics.

The distance between a pair of clusters, c_i and c_j, is computed by combining distances between the audio and video models as

$$d(c_i,c_j) = (2-\beta_{ij}) \quad d(a_i,a_j) \quad + \quad \beta_{ij} \quad d(v_i,v_j) \qquad (7)$$

where $d(a_i,a_j)$ and $d(v_i,v_j)$ are computed using Equation 2 and Equation 5, respectively and β_{ij} is a weighting term that determines the influence of video on the overall distance, calculated using

$$\beta_{ij} = min(\beta_i,\beta_j) \quad \text{where} \quad \beta_i = \sum r_{i,\gamma} \qquad (8)$$

Equation 8 requires some explanation. The eigen-blob model r_i for a cluster is sometimes fragmented over multiple persons. This happens because of consistent co-occurring motion such as hand movements of a person who takes notes when someone else is speaking. Such fragmentation incorrectly reduces the distance between video models of the involved persons, and negatively influences the clustering procedure. Let $r_{i,\gamma}$ be the connected component of r_i that represents the maximum fraction of r_i, i.e. it is the blob that captures the maximum fraction of the pdf. Then, β_i, which is the sum over $r_{i,\gamma}$ can be considered a fragmentation measure; β_i will be one if r_i is not fragmented, and low if r_i is severely fragmented. The weighting term β_{ij} represents the confidence in the computed video distance. If either of the two eigen-blob models is fragmented, β_{ij} will have a low value, reflecting lesser confidence in the localization and reducing the contribution of $d(v_i,v_j)$ to the inter-cluster distance.

Equation 7 is used to compute the pairwise distance between all of the intermediate clusters and the pair with the lowest distance is merged. GMMs for speech and eigen-blob models for video are now built from the merged cluster and the iterative procedure continues till a stopping criterion is met.

Ideally, the stopping criterion should terminate the iterations when the number of final clusters is equal to the number of speakers. In previous work dealing with audio-only diarization, the ΔBIC criterion has been extensively used [15] as the stopping criterion. In our experiments we found that using only the ΔBIC, tends to result in lesser clusters than the number of participants. This occurs when clusters from two speakers are incorrectly merged if the speaker's have similar vocal characteristics or if the clusters are impure, i.e. they contain speech from more than one speaker. Since eigen-blob models built from the intermediate clusters localize the speaker, eigen-models for different speakers lie on different regions of the image. Thus, if the eigen-blob models for two clusters do not overlap, the clusters are most likely from different speakers. Taking video into account, the stopping criterion terminates the iterations if either $r_i \cap r_j = \emptyset$, or $\Delta\text{BIC}(a_i,a_j) > 0$. This combined use of audio and video leads to a more robust stopping criterion.

V. Results

The proposed audio and audio-visual speaker diarization and localization approaches are tested on sixteen meetings from the NIST pilot meeting room corpus [9]. For each meeting, four camera feeds are available (one camera on each wall of the room). The videos have a spatial resolution of 720 x 480 sampled at 29.97 Hz. There are two audio channels packaged with each video; one is a gain-normalized mix of the head microphones worn by the participants, and the second is a gain-normalized mix of distant microphones placed on the central table and the wall. The audio data is sampled at 44 kHz and has a resolution of 16 bits per sample. Eight audio-visual pairings are considered for each meeting by pairing each of the four cameras with each of the two audio channels, resulting in 128 (16 x 8) meeting clips. From each clip, the first 30 seconds are discarded, and the next 10 minutes are chosen resulting in approximately 21 hours of data.

In the meetings, participants are seated around a central table and interact casually. Depending on the type of the meeting, the participants discuss a given topic, plan events, play games or attend presentations. From time to time, participants may take notes, stretch, and sip drinks. The audio and video signals from these meetings are quite complex because the meetings are unscripted and of long durations. Since only a single camera view is considered at a time, most faces are non-frontal and sometimes participants are only partially visible. In some meetings, a participant may not be visible at all in a particular camera view. Similarly, the audio signal is complex, consisting of short utterances, frequent overlaps in speech, and non-speech sounds such as wheezing, laughing, coughing, etc. Additionally, in some of the meetings (5 and 9-12), participants leave their chairs to use the white-board or distribute materials. Sample images of four clips from two of the camera views are shown in Figure 4.

To quantify the localization performance, the ground-truth is defined by static boxes around each person. Eigen-blob localization outputs a dominant blob r_γ for each of the final clusters. This is a static region in the image which localizes the person in all meeting frames where the person spoke. The output of MI localization is the connected component of the MI image with the highest average MI and so this region varies from frame to frame. For a frame t, a hit occurs if more than 50% of the region output by a localization method for that frame $S(t)$, intersects with the ground-truth box around the speaker $B(t)$. Mathematically, a hit is defined as

$$h(t) = \left\{ \begin{array}{ll} 1 & \text{if} \quad |S(t) \cap B(t)| > 0.5|S(t)| \\ 0 & \text{otherwise} \end{array} \right\} \qquad (9)$$

where $|S(t)|$ is the number of pixels in the system output, and $|S(t) \cap B(t)|$ are the number of pixels in the overlap between the system output and the ground truth bounding box. For eigen-blob localization, $S(t) = r_{i,\gamma}$, the dominant blob of the eigen-model for the cluster (i) that contains frame t. For

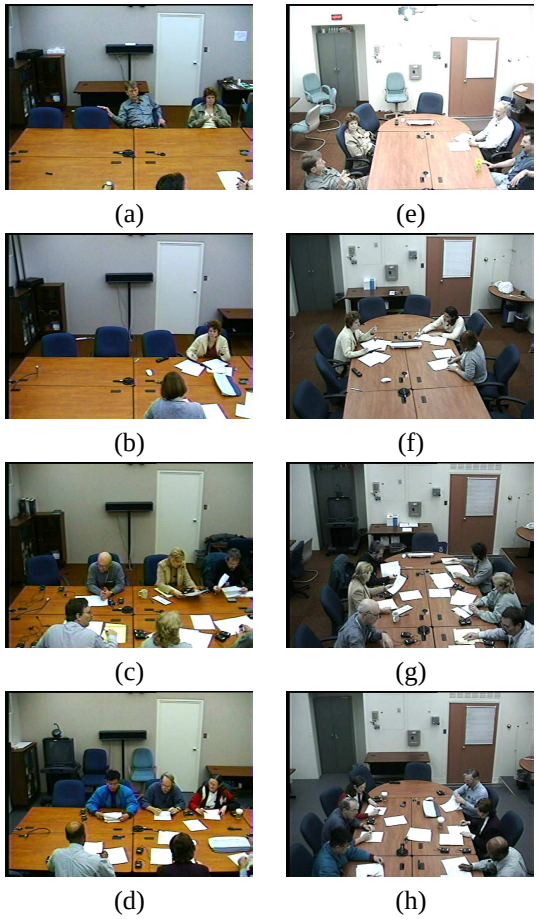

Fig. 4. Sample images from two views for four of the meetings. Images (a-d) are from a camera that captures the frontal and parietal view of the participants. Images (e-h) are from a camera that shows most participants in profile view. The other two other cameras provide similar views.

MI based localization, S(t) is the blob with the highest average MI in frame t.

Only non-overlapping speech frames where the speaker is visible (completely or partially) in a camera view are evaluated for localization. Representing the subset of frames over which localization is scored as T_e, the hit ratio is computed as the ratio of hits to the number of frames in T_e as

$$HitRatio = \sum_{t \in T_e} h(t)/|T_e| \tag{10}$$

The diarization performance is measured using the diarization error rate (DER) defined in [9]. Only non-overlapping speech frames are scored. To compute the DER, a one-one mapping is performed between the system clusters (final clusters) and the reference clusters (ground-truth), such that the mapping maximizes the total number of frames in agreement. The DER is then computed as

$$DER = E_{MISS} + E_{FA} + E_{spkr} \tag{11}$$

where E_{MISS} is the percentage of scored frames where speech is classified as silence and E_{FA} is the percentage of scored frames where silence is classified as speech. These errors occur due to imperfect speech/silence classification. E_{spkr} is the percentage of scored frames where speech from one speaker is incorrectly attributed to another speaker. Missed ATP boundaries, imperfect clustering and incorrect stopping contribute to this error.

A. Localization

Figures 5 and 6 illustrate the eigen-blob models for some of the final clusters in meetings 3 and 6. Meeting 3 is a planning meeting with frequent note taking activity, while meeting 6 is a card game scenario with participants frequently reaching out to the center of the table to pick and drop cards. Figures 5 (a-d) show the eigen-blob models (r_i) for the four final clusters of meeting 3 and (e-h) show the models for four of the six clusters of meeting 6 in the first camera view. The models lie only on the speakers (green boxes) for the two easy cases (b) and (g) where the speaker is frontal and also for the difficult cases (a), (c) and (e) where the speakers are non-frontal or partially hidden. In (d), (f) and (h), we see that fragments of the models lie on non-speakers (red boxes). However, for (f) and (h), the dominant blob, ($r_{i,\gamma}$) still lies on the speaker.

Figure 6 shows the eigen-blob models for the same data in the second camera view where the participants appear in profile view. The models lie only on the speaker in most cases, but blobs lie on non-speakers in (c), (d) and (f). In general, we find that when r_i is not fragmented, it usually lies on the correct speaker and even when r_i is fragmented, the dominant blob ($r_{i,\gamma}$) still localizes the speaker correctly. However, when a non-speaker exhibits consistent motion that exceeds the speaker's motion, $r_{i,\gamma}$ incorrectly localizes the non-speaker, as occurs in 5(d) and 6(d).

Figure 7 compares the localization performances of the MI and eigen-blob methods using the $HitRatio$ metric defined in Equation 10. Each audio channel is paired with the four cameras and the localization result is presented as the mean of the four $HitRatios$ with error bars indicating the maximum and minimum of the four values. The localization results for the two audio channels are shown separately for the two localization methods resulting in four bars per meeting. For each meeting, the subset of frames T_e over which the $HitRatio$ is computed may differ if all speakers are not visible in all camera views.

From Figure 7 we observe that the eigen-blob localization procedure works well in most meetings but performs poorly on a subset of meetings (5, 9-12). As mentioned earlier, these meetings violate the assumption that the participants stay seated, which leads to poor localization results. This is because the eigen-blob models are split between the true speaker and a moving participant. The large motion magnitude generated by a moving person, causes the r_γ blob to localize the moving person instead of the speaker. Incidentally, these are the only meetings where MI based localization performs better. This is because the MI is computed over short time windows and hence unaffected by a change in speaker location.

The average $HitRatio$ across the dataset for the eigen-blob localization method is 65.24% and 62.04% for channel 1 and 2, respectively which is substantially higher than the 51.3% and 49.54% obtained using MI. If the five meetings (5, 9-12) are dropped, the difference is even more pronounced,

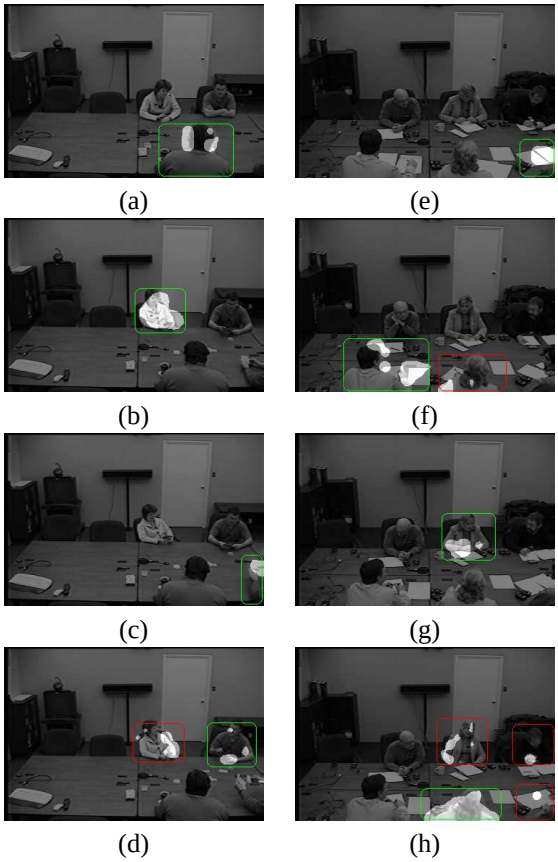

Fig. 5. Localization of speakers in the first camera view. (a)-(d) shows localization of the four speakers from meeting 3 and (e)-(h) shows localization of four of the six speakers from meeting 6. A green box shows the true speaker and a red box indicates where a non-speaker is covered by the model. In (d), (f), and (h), fragments of the eigen-blob model lie on the non-speakers. However, except in (d), the dominant blob correctly localizes the speaker.

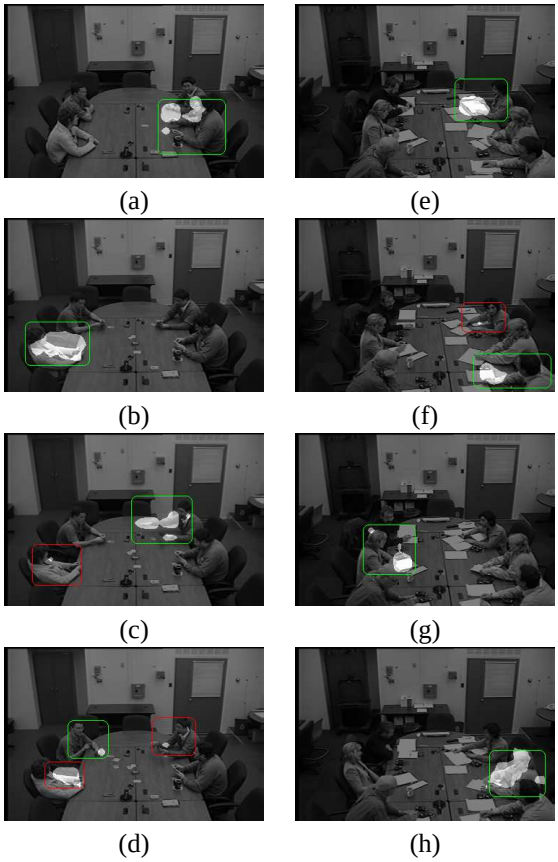

Fig. 6. Localization of speakers in the second camera view. (a)-(d) shows localization of four persons from meeting 3 and (e)-(h) shows localization of four of the six participants of meeting 6. In (c), (d), and (h), fragments of the eigen-blob model lie on the non-speakers.

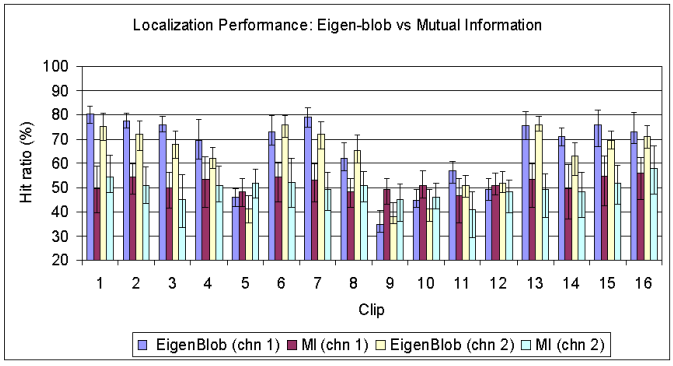

Fig. 7. Localization performance using the eigen-blob and Mutual Information (MI) based methods. The r_γ blob is considered as the system output for the eigen-blob method and the region with the highest average MI is considered as the system output for the MI method. A hit occurs if more than 50% of the system output overlaps with the speaker's true location.

with eigen-blob localization yielding 73.8% and 69.97% for channels 1 and 2, respectively compared to MI's 52.34% and 50.9% for the respective channels. Comparing across channels, we observe that the localization methods tend to performs better with channel 1 as it's speech quality is better than that of channel 2. We also see that the variation of performance across cameras is much lower for the eigen-blob method than the MI method. The MI method performs better when the dominant speaker faces the camera whereas the eigen-blob method is much more invariant to change in camera views.

Errors in eigen-blob localization stem from two sources: one, an intermediate cluster may contain speech from other speakers and those frames will be marked with the location of the dominant speaker. Two, non-speakers that exhibit continuous motion over a long duration (swiveling on a chair throughout the meeting), will cause fragments of the eigen-blob model to lie on their location. If the non-speaker motion is consistent and of larger magnitude than the speaker's motion, the r_γ blob incorrectly localizes the non-speaker.

Since no audio clustering is performed in the MI based methods, the method is not affected by diarization errors. Errors occur when a non-speaker's movements show stronger association with the audio signal, which occurs when a listener exhibits significant motion for short durations. MI incorrectly localizes such movements incurring a drop in performance. The situation worsens when the speaker is facing away from the camera - as motion from the speaker is less visible and hence easily overwhelmed by spurious background motion.

B. Diarization

In this subsection, we quantify the influence of video on diarization. The framework for audio-only diarization is

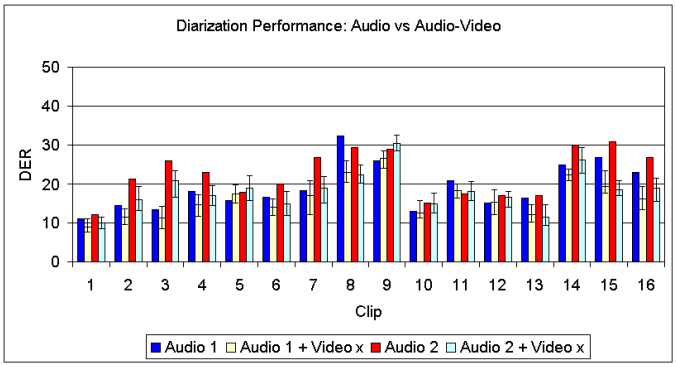

Fig. 8. Comparison of diarization performance using audio-only and audio-video information. For most meetings, the incorporation of video information results in lower *DER*. Meetings 5, 9-12 do not gain much improvement by incorporating video because of poor localization in these meetings.

similar to that of audio-visual diarization, except that no video features were used. For each meeting, two audio-only diarization results are obtained - one for each channel. Similar to localization, four diarization results are obtained for each audio channel, by combining each of the four cameras individually with the audio channel. The results are presented as the mean of the four results, with error bars indicating the maximum and minimum of the four results.

Figure 8 compares the performance of the diarization scheme when using only audio to that when using both audio and video. The average *DERs* of audio-only diarization are 19.15% and 22.54% for channels 1 and 2, respectively. The incorporation of localization results in average *DERs* of 16.27% and 18.42% which corresponds to relative improvements of 15.0% and 18.02%, respectively. However, video does not always improve diarization performance as seen from meetings 5 and 9-12. This is because some participants leave their seats for short durations leading to fragmented video models and resulting in a low value of β when computing inter-cluster distances using Equation 7. If these meetings are eliminated, the average *DERs* for channel 1 and 2 are 19.62% and 24.01% for audio-only diarization and 15.47% and 17.8% for audio-video diarization. This represents relative improvements of 21.16% and 25.87%, respectively.

Figure 8 reveals that the average *DER* for channel 2 is higher than that for channel 1. This is expected, since channel 1 is obtained from head microphones and thus has better quality than channel 2 which is recorded from distant microphones. A similar pattern is found for audio-visual diarization, since the localization process is essentially guided by the audio.

Errors in the audio-only diarization stem from the speech/silence detector, missed ATP boundaries and incorrect clustering. In addition to these, poor localization contributes to errors in audio-visual diarization. Poor localization may place fragments of r_i on a non-speaker, incorrectly reducing the video distance between models belonging to different speakers. In the extreme case, where the speaker is not visible in a particular camera - this situation is inevitable. However, in such cases, the eigen-blob model is split almost evenly amongst the other participants. This results in a low value of β for r_i, reducing the contribution of the video to the inter-cluster distance and preventing poor localization from adversely affecting diarization.

Figure 8 shows that audio-video diarization outperform audio-only diarization. Since video is incorporated at two stages in our system - for segmenting the joint feature stream into ATPs and during the iterative clustering stage, we conducted experiments to evaluate the influence of video in each of these stages which are tabulated in Table I.

TABLE I
IMPACT OF VIDEO ON AUDIO DIARIZATION.

Use of Video		Average DER	
ATP Segmentation	Iterative Clustering	Audio 1	Audio 2
-	-	19.15	22.54
-	√	17.16	19.31
√	-	18.46	22.0
√	√	16.27	18.42

Table I indicates that the major benefit of incorporating video comes from it's participation in the iterative clustering stage. This is expected as the intermediate clusters contain far more data than the ATPs and so incorrect clustering impacts the performance much more than missed ATP boundaries. However, for a couple of meetings we found that incorrect segmentation lead to impure ATPs that were subsequently clustered incorrectly, and significantly impacted the DER. Thus, rather than provide performance gains, the use of video for ATP segmentation imparts robustness to the system.

VI. SUMMARY AND CONCLUSIONS

This paper presents a novel approach to perform speaker localization and diarization in meetings recorded by a single camera and a single microphone. Previous approaches dealing with joint audio-visual analysis in this scenario seek correlations between the two spaces. These solutions assume that the audio and video signals are instantaneously correlated, but as demonstrated in this work, the assumption does not hold in the meeting domain. In the proposed approach, instead of formulating the problem as finding correlations across spaces, clustering is performed in individual spaces. The association of clusters across spaces is based on the assumption that speech and body movement co-occur.

The approach is evaluated on a substantially large dataset (21 hours) of unscripted real meetings. The dataset is obtained by pairing two audio channels of different sound qualities with four different camera views. Localization results on this challenging dataset find that the eigen-blob based method outperforms the MI based method by about 40% (relative). In addition, the eigen-blob based localization is less sensitive to changes in camera view.

The novelty of the diarization process is in its use of motion information from the entire body, rather than just the face. Results obtained by incorporating video information into the clustering process leads to a relative improvement of about 16% over that of using audio alone. These are encouraging results given the nature of the meetings and the video quality.

The system performance will improve if each participants face was visible, and the lips were tracked. However, these results show that there are cues in the video beyond the face

that tell of speech activity. This work exploits such video information on a global scale without relying on explicit face/head/hand detection and without assuming frontal faces. Also, since the approach does not require training or *a priori* information, it is readily adaptable to other domains.

REFERENCES

[1] Z. Barzelay and Y. Schechner, "Harmony in motion," in *Computer Vision and Pattern Recognition*, 2007, pp. 1–8.

[2] M. Ben, M. Betser, F. Bimbot, and G. Gravier, "Speaker diarization using bottom-up clustering based on a parameter-derived distance between adapted GMMs," in *International Conf. on Spoken Language Processing*, 2004, pp. 2329–2332.

[3] C. Busso, S. Hernanz, C.-W. Chu, S.-I. Kwon, S. Lee, P. Georgiou, I. Cohen, and S. Narayanan, "Smart room: Participant and speaker localization and identification," in *IEEE International Conf. on Acoustics, Speech and Signal Processing*, vol. 2, 2005, pp. 1117–1120.

[4] S. S. Cheng and H. M. Wang, "A sequential metric-based audio segmentation using Bayesian information criterion," in *Eurospeech*, 2003.

[5] R. Cutler and L. Davis, "Look who's talking: Speaker detection using video and audio correlation," in *IEEE International Conf. on Multimedia*, 2000, pp. 1589–1592.

[6] R. Cutler, Y. Rui, A. Gupta, J. Cadiz, I. Tashev, L. He, A. Colburn, Z. Zhang, Z. Liu, and S. Silverberg, "Distributed meetings: A meeting capture and broadcast system," in *ACM Multimedia*, 2002, pp. 503–512.

[7] J. Fiscus, N. Radde, J. Garofolo, A. Le, J. Ajot, and C. Laprun, "Rich transcription 2005 spring meeting recognition evaluation," Accessed: April 2007. [Online]. Available: www.nist.gov/speech/publications

[8] J. Fisher and T. Darrell, "Probabalistic models and informative subspaces for audiovisual correspondence," in *European Conf. on Computer Vision*, vol. 3, 2002, pp. 592–603.

[9] J. S. Garofolo, C. D. Laprun, M. Michel, V. M. Stanford, and E. Tabassi, "The NIST meeting room pilot corpus," in *International Conf. on Language Resources and Evaluation*, 2004.

[10] J. Hershey and J. Movellan, "Audio vision: Using audiovisual synchrony to locate sounds," in *Advances in Neural Information Processing Systems*, 1999, pp. 813–819.

[11] B. Kapralos, M. Jenkin, and E. Milios, "Audio-visual localization of multiple speakers in a video teleconferencing setting," *International Journal of Imaging Systems and Technologies*, vol. 13, pp. 95–105, 2003.

[12] E. Kidron and Y. Schechner, "Pixels that sound," in *Computer Vision and Pattern Recognition*, 2005, pp. 88–95.

[13] G. Monaci, O. D. Escoda, and P. Vandergheynst, "Analysis of multimodal sequences using geometric video representations," *Signal Processing*, vol. 86, pp. 3534 – 3548, 2006.

[14] H. J. Nock, G. Iyengar, and C. Neti, "Speaker localization using audio-visual synchrony: An empirical study," in *ACM International Conf. on Multimedia*, 2003, pp. 488–499.

[15] J. Pardo, X. Anguera, and C. Wooters, "Speaker diarization for multiple-distant-microphone meetings using several sources of information," *IEEE Transactions on Computers*, vol. 56, pp. 1212–1224, 2007.

[16] E. Patterson, S. Gurbuz, Z. Tufekci, and J. Gowdy, "Moving talker, speaker-independent feature study and baseline results using the CUAVE multimodal speech corpus," *Journal on Applied Signal Processing*, vol. 11, pp. 1189–1201, 2002.

[17] J. Rehg, K. Murphy, and P. Fieguth, "Vision-based speaker detection using Bayesian networks," in *Computer Vision and Pattern Recognition*, 1999, pp. 110–116.

[18] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital Signal Processing*, vol. 10, pp. 19–41, 2000.

[19] M. Sargin, O. Aran, A. Karpov, F. Ofli, Y. Yasinnik, S. Wilson, E. Erzin, Y. Yemez, and A. Tekalp, "Combined gesture-speech correlation analysis and speech driven gesture synthesis," in *IEEE International Conf. on Multimedia*, 2006, pp. 893–896.

[20] S. Sarkar and P. Soundararajan, "Supervised learning of large perceptual organization: Graph spectral partitioning and learning automata," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 5, pp. 504–525, 2000.

[21] M. Slaney and M. Covell, "Facesync: A linear operator for measuring synchronization of video facial images and audio tracks," in *Advances in Neural Information Processing Systems*, vol. 14, 2000, pp. 814–820.

[22] R. Stiefelhagen, K. Bernardin, R. Bowers, J. Garofolo, D. Mostefa, and P. Soundararajan, "The CLEAR 2006 evaluation," *Springer Lecture Notes in Computer Science*, no. 4122, pp. 1–44, 2006.

[23] S. Tranter and D. Reynolds, "An overview of automatic speaker diarization systems," *IEEE Transactions on audio speech and language processing*, vol. 14, no. 5, pp. 1557–1565, 2006.

[24] H. Vajaria, T. Islam, S. Sarkar, R. Sankar, and R. Kasturi, "Audio segmentation and speaker localization in meeting videos," in *International Conf. on Pattern Recognition*, vol. 2, 2006, pp. 1150–1153.

[25] L. Valbonesi, R. Ansari, D. McNeill, F. Quek, S. Duncan, K. E. McCullough, and R. Bryll, "Multimodal signal analysis of prosody and hand motion: Temporal correlation of speech and gestures," in *European Signal Processing Conf.*, 2002, pp. 1330–1345.

[26] C. Wooters, J. Fung, B. Peskin, and X. Anguera, "Towards robust speaker segmentation: The ICSI-SRI fall 2004 diarization system," in *Rich Transcription Workshop*, 2004.

Himanshu Vajaria Himanshu Vajaria received the MS degree in electrical engineering from the Pennsylvania State University in 2003 and a PhD in computer science and engineering from the University of South Florida in 2008. He received the best student paper award at ICPR 2006. His research interests include biometrics, image and video processing, computer vision, pattern recognition and remote signal processing. He is a student member of the IEEE and the IEEE Computer Society.

Sudeep Sarkar Sudeep Sarkar received the BTech degree in electrical engineering from the Indian Institute of Technology, Kanpur, India, in 1988 and the MS and PhD degrees in electrical engineering, on a University Presidential Fellowship, from Ohio State University, Columbus, in 1990 and 1993, respectively. Since 1993, he has been with the Computer Science and Engineering Department at the University of South Florida, Tampa, where he is currently a professor. His research interests include perceptual organization in single images and multiple image sequences, biometrics, gait recognition, color-texture analysis, and performance evaluation of vision systems. He has coauthored one book and coedited another book on perceptual organization. He is the recipient of the US National Science Foundation Faculty Early Career Development (CAREER) award in 1994, the University of South Florida (USF) Teaching Incentive Program Award for undergraduate teaching excellence in 1997, the Outstanding Undergraduate Teaching Award in 1998, and the Theodore and Venette Askounes-Ashford Distinguished Scholar Award in 2004. He served on the editorial boards for the IEEE Transactions on Pattern Analysis and Machine Intelligence (1999-2003) and Pattern Analysis and Applications Journal (2000-2001). He is currently serving on the editorial boards of the Pattern Recognition journal, IEEE Transactions on Systems, Man, and Cybernetics (Part-B), Image and Vision Computing, and IET Computer Vision. He is a senior member of the IEEE and the IEEE Computer Society.

Rangachar Kasturi Rangachar Kasturi received the BE (electrical) degree from Bangalore University, India, in 1968 and the MSEE and PhD degrees from Texas Tech University in 1980 and 1982, respectively. He was a professor of computer science and engineering and electrical engineering at the Pennsylvania State University during 1982-2003 and was a Fulbright Scholar during 1999. He has been elected to serve as the 2008 President of the IEEE Computer Society. He was the President of the International Association for Pattern Recognition (IAPR) during 2002-2004. He has served as the editor in chief of the IEEE Transactions on Pattern Analysis and Machine Intelligence and the Machine Vision and Applications journals. He has received the Penn State Engineering Society Premier Research Award and has been inducted into the Texas Tech Electrical Engineering Academy. His research interests are in computer vision and pattern recognition. He is an author of the textbook Machine Vision and has published numerous papers and research reference books. He has directed many research projects in document image analysis, video sequence analysis, and biometrics. In particular, he is directing a project that evaluates research progress in detection and tracking of faces, people, text, and vehicles in video sequences. He is a fellow of the IEEE and a fellow of the International Association for Pattern Recognition (IAPR).