REAL-TIME MONITORING OF PARTICIPANTS' INTERACTION IN A MEETING USING AUDIO-VISUAL SENSORS

Carlos Busso, Panayiotis G. Georgiou and Shrikanth S. Narayanan

Speech Analysis and Interpretation Laboratory (SAIL)
Viterbi School of Engineering,
University of Southern California,
Los Angeles, CA 90089, USA

ABSTRACT

Intelligent environments equipped with audio-visual sensors provide suitable means for automatically monitoring and tracking the behavior, strategies and engagement of the participants in multiperson meetings. In this paper, high-level features are calculated from active speaker segmentations, automatically annotated by our *smart room* system, to infer the interaction dynamics between the participants. These features include the number and the average duration of each turn, statistics of turn-taking such as time as active speaker, and turn-taking transition patterns between participants. The results show that it is possible to accurately estimate in real-time not only the flow of the interaction, but also how dominant and engaged each participant was during the discussion. These high-level features, which cannot be inferred from any of the individual modalities by themselves, can be useful for summarization, classification, retrieval and (after action) analysis of meetings.

Index Terms— Intelligent sensors, Smart Room, Human Interaction, Human Factors

1. INTRODUCTION

Group meetings are a crucial part of planning and organization for any institution. If meetings could be automatically recorded and annotated, it would be possible to retrieve important information that transpired, which could prove helpful in analyzing teamwork and collaboration strategies, and contributing to productivity. Towards that direction, new technologies in sensing, tracking, storage, and retrieval are offering exciting and challenging applications for human interaction sensing and human centered computing. Example realms in which monitoring human interaction is very useful are retrieval [1], summarization [2] and classification [3, 4, 5] of meetings. While much of the prior work has focused on content analysis (such as speech transcription), there is increasing interest in expanding it to include meta-information such as affect, speaker dynamics, etc. Since it is expected that the number of meetings being archived will rapidly increase, especially given interactions across the globe, automatic annotations of meta-information of human interaction will play an important role in efficient and intuitive searching of specific portions and aspects of the meeting. This leads to even more interest into novel methods for robustly monitoring and measuring human interactions. In this context, *Smart rooms* equipped with non-intrusive multimodal sensors provide a suitable platform for automatically inferring meta-information from the participants in meeting and control room type environments. This paper focuses on meta-analysis of certain aspects of group meetings from audio-visual information obtained in a smart room.

Recent efforts to infer high-level information from meetings include [3], where the authors evaluated the use of interactive features extracted from manual annotations to classify the meeting (e.g. discussion, presentation) and each participant's role (e.g. presenter, listener). A similar goal was pursued in [4] and [6], in which HMM-based approaches were implemented to detect meeting actions, using features extracted from the audio-visual sensors. The influence between participants was studied in [7]. They learned behavioral models to predict who is most likely to take the next turn. In [8], the author tracked the level of dominance of the participants. In most of these works, manual annotations were used to extract the features, which make these approaches not directly applicable for real-time applications.

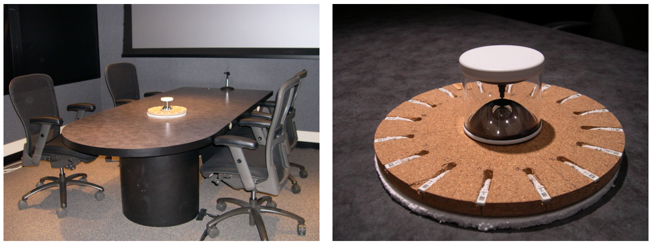


Fig. 1. Smart Room. The left figure shows the smart room. The right figure shows the microphone array and the omnidirectional camera.

In our previous work [9], we presented the *Smart room* environment being developed at the *University of Southern California* (USC). This conference room employs microphones and cameras for activity sensing. The sensors were appropriately fused to estimate the spatial position of the users, detect speaker activity and determine the speaker's identity. Building upon that work, this paper evaluates the use of high-level information extracted from this environment to monitor the participants' behavior. For each subject in the room, the number of turns, the percentage of time during active speaking, the average length of the turns, and the turn-taking transition matrix between participants were measured. The results show that these high-level features provide information about the flow of the interaction that cannot be accurately inferred from any of the individual sensors. They also provide dynamic estimation of participant engagement/involvement during the meeting. Since these features are estimated from automatic speaker segmentations, the proposed system can monitor human interaction in real-time, making feasible the interesting applications mentioned before.

2. SMART ROOM

The current setting of our smart room is very similar to the one presented in [9]. It consists of a microphone array with 16 omnidirectional microphones, four firewire CCD cameras near the corners of the ceiling, and one full-circle omnidirectional camera (Fig. 1). In addition, a directional microphone at 16KHz was added at one side of the table. This section briefly summarizes the purpose of each modality, and the methodology used to process the captured data. More information about each modality can be found in [9].

1-4244-0728-1/07/$20.00 ©2007 IEEE

2.1. Audio sensors

The microphone array was used for acoustic source localization. The approach is based on the *Time Difference of Arrival* (TDOA) of the sound to the various microphones. The geometric inference of the source location is calculated from this TDOA. First, pair-wise delays are estimated between all the microphones ((16 × 15)/2 =120) [10]. These delays are subsequently projected as angles into a single axes system. This results in 240 *direction-of-arrival* (DOA) estimates, half of them stemming from the front-to-back confusion of the microphone pairs. The density of these estimates provides a mode that corresponds to the correct DOA of sound.

In our previous setting [9], the microphone array was placed at one side of the table, and as a consequence, the useful aperture of the array was small. In the current study, the microphone array was placed in the middle of the table in a 2-D structure (see Fig. 1). Although this configuration does not provide depth information, the wide DOA range allows the microphone array to separate the participants' speech with higher accuracy (approximately from 70% to 85%). The depth information is redundant in this setting due to the 4-camera system's complementary measurements, which compensate amply for the inaccurate range information that an array can provide for a far-field source.

The extra directional microphone was used for supervised speaker identification. A *Gaussian Mixture Model* (GMM) with 16 mixtures was used to recognize the speaker identity. From the acoustic signal, 12 *Mel Frequency Cepstral Coefficients* (MFCC) were estimated to train the model, using standard methods such as *Expectation-Maximization* (EM) and *Maximum Likelihood* (ML). In addition, a background model was added to detect periods of silence. The speaker detection is calculated every 1 second.

2.2. Video sensors

Four CCD ceiling cameras are used to detect and track the spatial location of the speakers. First, moving regions in the scene are segmented by comparing each frame with a Gaussian background-learning model. After compensating for shadows, silhouettes of the moving object in the room are created. Then, the detected silhouettes across the views are fused to estimate the 3-D visual hulls of the people in the room [11]. Following that, a polygon approximation is fit to each hull, and a polyhedral representation is computed directly from these polygons [12]. Finally, the polygon surface is randomly sampled and a height map is constructed. The local maxima of the height map are detected and considered as heads of the participants. Thresholds are used to remove small areas such as chairs and papers.

A full-circle omidirectional camera is used to detect the angles of the participants' faces. The acquired images are the result of the projection of the surrounding scene into a hemisphere, which are then unrolled and projected back into cylinders (see Fig. 2). To detect the foreground region, Gaussian background-learning models are also used to compare the color distribution of new frames to detect moving blobs prior to capturing the faces. Morphological operators are used to group detected pixels into foreground regions, and small regions are eliminated. In these moving regions, face detection based on Haar-like features was implemented, using Intel's open source computer vision library [13]. Notice that the color histogram of the detected regions is normalized beforehand to accurately detect faces under low light level conditions.

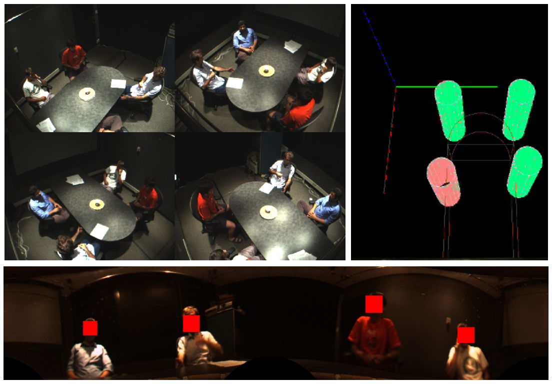


Fig. 2. Speaker localization system. See Sec 2.2 and 3.1 for details.

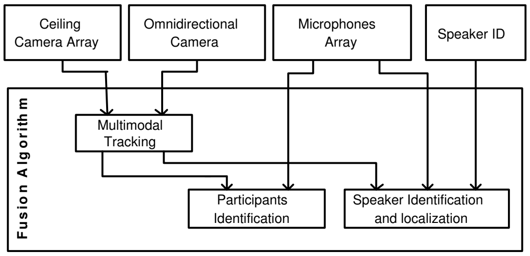


Fig. 3. The system is distributed running over TCP, with information exchange as depicted above.

3. MULTIMODAL FUSION

3.1. Participant Localization and Identification

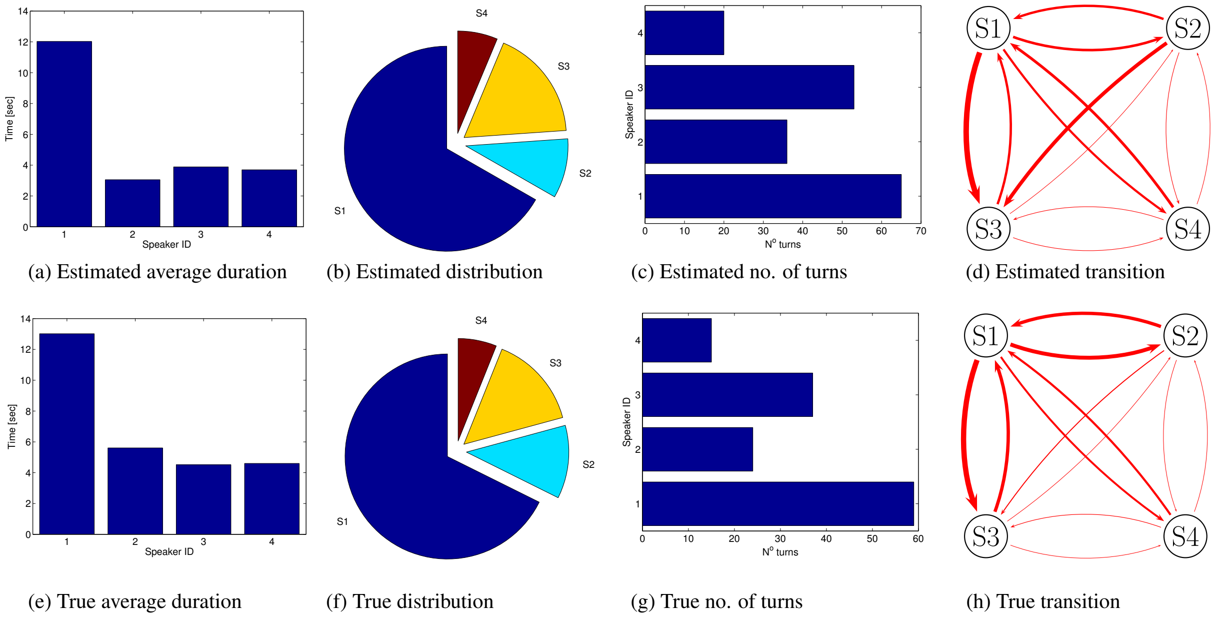
Figure 3 describes the fusion technique used to locate the participants, to recognize their identities, and to infer the active speaker over time. The real-time system is distributed running over TCP. Each modality sends information to the fusion algorithm that makes the decision every 1 second, which is the slower frame rate of the modalities (speaker ID).

The tracking algorithm makes use of the visual modalities. Participants are sequentially identified entering the room if consistent measures are detected by the ceiling cameras. The angles of the detected faces are used to correct the true participants' locations. Since errors in both modalities are independent, the tracking algorithm is quite robust.

After the participants' locations are estimated, the microphone array and the speaker ID are used to detect the active speaker. In the angle domain, the participants are modeled with a Gaussian distribution, with a mean estimated from the participants' position (in angle), and a constant variance empirically chosen. This is referred to as $P(S_i|X_{MA})$, the probability that the active speaker is participant i, given the microphone array. The speaker ID is also used to infer the active speaker, $P(S_i|X_{SID})$. Since the seating arrangement (L) is unknown, correlation with physical constraints, such as that one participant can only be at one point in space, is computed to estimate L. Finally, the active speaker is obtained by multiplying $P(S_i|X_{MA}) \cdot P(S_i|X_{SID})$, under the assumption of independence. By using these two modalities together, both the active speaker and the participant identities are estimated.

3.2. Performance Evaluation

Three 20-minute meetings with four participants were recorded and processed with the system. Since the participants were asked to speak as naturally as possible, the overlap between speakers was between 7% and 15%, which make this a challenging database. The meetings were segmented by hand to provide ground truth and compared to the results given by the fusion algorithm. Two criteria were defined: *strong decision*, correct if the speaker was the most active



(a) Estimated average duration (b) Estimated distribution (c) Estimated no. of turns (d) Estimated transition

(e) True average duration (f) True distribution (g) True no. of turns (h) True transition

Fig. 4. High-level group interaction measures estimated from automatic (a-d) and manual (e-h) speaker segmentation.

speaker during the time interval; and *weak decision*, correct if the speaker was active during the time interval. For instance, if the detected participant spoke during the interval, but someone else spoke more than him/her, then the results is considered correct only under the *weak criterion*. The results are presented in Table 1. Compared with our previous results [9], the performance accuracy has significantly increased (from 75% to 85%, see row D), which may be explained by the new configuration and redundancy exploitation of the microphone array.

		Session	Strong Decision	Weak Decision
A	Speaker ID (GMM based)	1	66.13%	73.28%
		2	61.27%	68.51%
		3	60.10%	67.85%
B	Microphone Array + Video	1	81.26%	86.02%
		2	85.41%	92.86%
		3	83.03%	89.62%
C	Microphone Array + Video + Speaker ID (assumes known seating arrangement L)	1	81.55%	88.42%
		2	85.60%	93.56%
		3	82.49%	90.32%
D	Microphone Array + Video + Speaker ID (Participant location (L) learned through data)	1	80.37%	87.34%
		2	78.77%	87.26%
		3	82.49%	90.24%
E	Seating arrangement automatically learned through data (L)	1	87.78	
		2	74.60	
		3	97.14	

Table 1. All of the above results are obtained in real time, and include the whole length of the meeting, with *no* time given for initial convergence. **A**: Speaker ID as obtained purely from the speech signal using a GMM; **B**: Localization obtained by the two visual information channels and the microphone array; **C**: Speaker Identification & Localization based on all information channels. Assumes perfect knowledge of L, the seating arrangement of the participants; **D**: As C, but the mapping of speaker-location, L, is continuously estimated from the data; **E**: Speaker Location mapping, L.

4. PARTICIPANT INTERACTION

In this section, high-level features derived from the automatic segmentation provided by the fusion algorithm are used to infer how people interact. For each participant, we calculated the number of turns, the average time duration of each turn, the amount of time used as active speaker, and the transition matrix depicting turn-taking between participants. Figures 4 (a-d) show the results for meeting 3. For reference and evaluation, Figures 4 (e-h) show the ground-truth for the same data obtained through human annotation.

Interesting observations can be inferred from these high-level features. The dominance of a participant is closely related to the distribution of time as the active speaker and the number of turns taken. [8]. We observe that subject 1 spoke more than 65% of the time, which suggests that he was most probably leading the discussion. This subject also presented the longest average duration for each turn, which suggests that his strategy was to present, elaborate and support his ideas. In contrast, the average duration for subject 3, who had the second largest number of turns, was only about 4 seconds, which reveals that he contributed with shorter sentences to support or contradict current ideas. These interpretations agree with previous work which suggests that discussions are characterized by the mediator taking long turns and the rest of the participants taking many short turn to show agreement (e.g. “uh-huh”) [14].

The transition matrix between participants provides further information about the flow of the interaction and the turn-taking patterns. By annotating the transition between speakers, a rough estimation can be inferred about who the speaker was addressing. To evaluate this hypothesis, we manually annotated whether the subject was speaking to all the participants or only to one of them. Table 2 compares the ground-truth annotations with the results provided by the transition matrix. As can be observed, the transition matrix provides a good first approximation to identifying the interlocutor dynamics.

	Hand-based addressee Annotation				Turn taking Transition Matrix			
	Sp1	Sp2	Sp3	Sp4	Sp1	Sp2	Sp3	Sp4
Sp1	0.00	0.31	0.44	0.25	0.03	0.34	0.46	0.17
Sp2	0.72	0.00	0.21	0.07	0.74	0.04	0.22	0.00
Sp3	0.69	0.18	0.00	0.13	0.76	0.08	0.05	0.11
Sp4	0.50	0.23	0.28	0.00	0.73	0.00	0.20	0.07

Table 2. Comparison between the hand-based addressee annotations and turn-taking transition matrix.

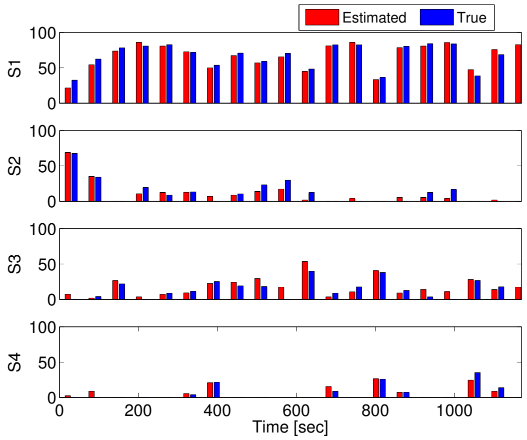


Fig. 5. Dynamic behavior of speakers' activeness over time.

In Figures 4 (d) and (h), the width of the arrows increases with the number of the times that one subject spoke after the other. Figure 4 (d) reveals that the discussion was mainly between subjects 1 and 3. In the presence of information such as context or change in affective states of the participants, this transition matrix could be extremely useful in determining coalition and rivalry between participants. Such information can in turn inform decision making efficacy in organizational communication.

The same high-level features can be estimated in small windows over time to analyze the dynamic behavior of the participants' interaction. Figure 5 shows the percentage of the time that each speaker was active in one minute windows. The figure also compares the results derived from automatic and manual annotations. A measure of the participants' engagement can be inferred from this dynamic feature. In this example, the figure shows that subject 4 only occasionally contributed in the discussion, suggesting that he was not engaged. Conversely, subjects 1, 2 and 3 participated in the discussion during the entire session. For a more reliable estimation of participants' engagement, recognition of gestures such as body posture and head orientation could be added to our current system. Again, these features can be useful for enriching meeting information retrieval. Knowing the segment of the meeting when a specific participant was speaking can help target the intended search. Such information can also be useful as training tool for improving participant skills during discussions. In posthoc analysis, for instance, after showing a meeting summary report to the participants, subject 1 declared that he was not aware of how dominant he was during the meeting. In the future, his strategies and style may change toward a more productive discussion in which everyone contributes with ideas.

Notice that these high-level features calculated from automatic and manual annotations do not significantly differ, which indicates that the fusion algorithm provides quite accurate active speaker segmentation. Therefore, the proposed system can be used in real-time monitoring of human interaction.

5. CONCLUSIONS

This paper evaluated the use of high-level information derived from automatic speaker segmentations, estimated by our *smart room* system, to infer how the participants interacted in a meeting. The results show that, with these features, it is possible to infer not only the flow of the discussion, but also the dominance and level of engagement of each participant. This information that cannot be accurately derived from any of the sensor modalities by themselves, is important for many applications such as summarization, classification, retrieval, and analysis of meetings.

Our ongoing research is focused on improving our *Smart room* system with the long-term goal of understanding how human beings communicate, especially in multiparty interactions. We are working to improve our tracking and fusion algorithms to have more reliable and robust active speaker localization. We are also directing our research efforts towards gesture recognition from the visual modalities, as well as spoken language processing, that can provide further detailed measures of participant emotions, awareness, and engagement.

Acknowledgment

This research was supported in part by funds from the NSF (through the Integrated Media Systems Center, a National Science Foundation Engineering Research Center, Cooperative Agreement No. EEC-9529152 and a CAREER award), the Department of the Army and a MURI award from ONR. Any opinions, findings and conclusions or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of the funding agencies. Thanks go to Tom Murray, Samuel Kim, Viktor Rozgic, Kyu Jeong Han, Chi-Wei Chu and Scott Millward.

6. REFERENCES

[1] A. Jaimes, K. Omura, T. Nagamine, and K. Hirata, "Memory cues for meeting video retrieval," in *1st ACM workshop on Continuous archival and retrieval of personal experiences (CARPE 2004)*, 2004, pp. 74– 85.

[2] I. Mikić, K. Huang, and M. Trivedi, "Activity monitoring and summarization for an intelligent meeting room," in *IEEE Workshop on Human Motion*, Austin, TX, USA, December 2000, p. 107 112.

[3] S. Banerjee and A.I. Rudnicky, "Using simple speech based features to detect the state of roles of the meeting participants," in *8th International Conference on Spoken Language Processing (ICSLP 04)*, Jeju Island, Korea, 2004.

[4] I. McCowan, D. Gatica-Perez, S. Bengio, G. Lathoud, M. Barnard, and D. Zhang, "Automatic analysis of multimodal group actions in meetings," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 3, pp. 305– 317, March 2005.

[5] S. Reiter, S. Schreiber, and G. Rigoll, "Multimodal meeting analysis by segmentation and classification of meeting events based on a higher level semantic approach," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2005)*, Philadelphia, PA, USA, March 2005, vol. 2, pp. 161– 164.

[6] D. Zhang, D. Gatica-Perez, S. Bengio, and I. McCowan, "Modeling individual and group actions in meetings with layered hmms," *IEEE Transactions on Multimedia*, vol. 8, no. 3, pp. 509– 520, June 2006.

[7] S. Basu, T. Choudhury, B. Clarkson, and A. Pentland, "Towards measuring human interactions in conversational settings," in *IEEE Int. Workshop Cues in Communication*, Kauai, HI, USA, December 2001.

[8] R.J. Rienks and D.K.J. Heylen, "Automatic dominance detection in meetings using easily obtainable features," in *Workshop on Multimodal Interaction and Related Machine Learning Algorithms*, Endinburgh, Scotland, October 2006, pp. 76– 86.

[9] C. Busso, S. Hernanz, C.W. Chu, S. Kwon, S. Lee, P.G. Georgiou, I. Cohen, and S. Narayanan, "Smart Room: Participant and speaker localization and identification," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2005)*, Philadelphia, PA, USA, March 2005, vol. 2, pp. 1117 – 1120.

[10] P. G. Georgiou, P. Tsakalides, and C. Kyriakakis, "Alpha-stable modeling of noise and robust time-delay estimation in the presence of impulsive noise," *IEEE Transactions on Multimedia*, vol. 1, no. 3, pp. 291–301, September 1999.

[11] A. Laurentini, "The visual hull concept for silhouette-based image understanding," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 16, no. 2, pp. 150162, Feb 1994.

[12] W. Matusik, C. Buehler, and L. McMillan, "Polyhedral visual hulls for real-time rendering," in *In Proceedings of Eurographics Workshop on Rendering*, 2001.

[13] A. Kuranov, R. Leinhart, and V. Pisarevsky, "An empirical analysis of boosting algorithms for rapid objects with an extended set of Haar-like features," in *Intel Technical Report MRL-TR-July02-01*, 2002.

[14] S. Burger, V. MacLaren, and H. Yu, "The ISL meeting corpus: The impact of meeting type on speech style," in *International Conference on Spoken Language (ICSLP)*, Denver,CO, USA, September 2002.