Associating Audio-Visual Activity Cues in a Dominance Estimation Framework

Hayley Hung[1] Yan Huang[2,3] Chuohao Yeo[3] Daniel Gatica-Perez[1,4]

[1] IDIAP Research Institute
Martigny, Switzerland
hhung@idiap.ch

[2] International Computer Science Institute (ICSI)
Berkeley, USA
yan@icsi.berkeley.edu

[3] University of California
Berkeley, USA
zuohao@eecs.berkeley.edu

[4] Ecole Polytechnique Federale de Lausanne (EPFL)
Switzerland
gatica@idiap.ch

ERRATUM: Please note after the time of publishing, some errors in the results were found. The results contained in this copy of the paper have been duly corrected.

Abstract

We address the problem of both estimating the dominant person in a meeting from a single audio source and identifying them visually in a multi-camera setting. We use a speaker diarization algorithm to perform speaker segmentation and clustering, representing when they spoke. Using a greedy ordered audio-visual association algorithm, we investigate using the speaker clusters to find the corresponding person in one of the video channels. The difficulty of the problem is that firstly the speaker diarization output is noisy (e.g. for participants who speak little) and often produces an unequal number of clusters to true participants. Secondly, personal visual activity from natural upper torso motion, which can include highly deformable pose changes and perspective distortion, is computed through computationally efficient coarse features. Our results using almost 2 hours of audio-visual data from 4-participant meetings show a strong correlation between the estimated speaker diarization and visual activity features, enabling the identification of the most dominant person as a pair of audio-visual channels.

1. Introduction

In meetings, groups of people gather to discuss and/or complete a task. Through this conversational process, it is natural for members to try to establish hierarchy in the group, even if the members are unacquainted [15]. One way that this can be observed is through dominant behaviour. Dominance has been studied in social psychology for several decades and has been described by Dunbar and Burgoon, as "...necessarily manifest. It refers to context and relationship-dependent interactional patterns in which one actor's assertion of control is met by acquiescence from another" [4] (p. 208) . It is often used synonymously with influence and power but social psychologists differentiate them by describing power as the "...capacity to produce intended effects, and in particular, the ability to influence the behavior of another person" [4] (p. 208). Importantly, in terms of inferring dominance through observable non-verbal cues, Schmid Mast [11] found, through a meta-analysis of several decades of literature, that dominance could be inferred through speaking time.

To our knowledge, Basu et al. [1] were the first to consider how influence could be automatically estimated. They modelled group interactions with the Influence Model (IM) with Markov chains where the transitions were affected by the influence that one participant could exert on another. Zhang et al. [18] expanded the idea further by suggesting the team-player influence model (TPIM), a two-layer dynamic Bayesian network which could model the influence of individuals on a group and vice versa. Rienks et al. [14] compared the TPIM with another method of estimating dominance using support vector machines . Otsuka et al. [13] also addressed influence by using an estimate of the visual focus of attention of each participant. We showed that simple speaking activity features automatically extracted from audio out-perform visual activity features extracted from video [7, 10]. In addition, as a single feature, the total speaking time of each participant is the best indicator of dominance. However there is still a need to automatically identify the most dominant person visually.

There has been much research in the area of audio-visual speaker association [9, 8, 16, 17] in which the scenario under investigation is that of spatio-temporally identifying one of two possible speakers in each video. Some of this work [9, 8, 16] has used very simple audio-visual data lasting between 2-20s where it is assumed that only one speaker utters a phrase. In some cases, the task is made more complicated by making one of the two people mouth the part of the phrase while the other utters them [16]. The limits of these works [9, 8, 16] are that they rely on video with near-frontal faces, the speakers do not move naturally, and the matching of audio and video signals are performed on very short time segments. In addition, there is only one speaker during each of the test sequences. Vajaria et al. [17] used more complex data from 4 five-minute long video clips, where they assumed that visual motion from a speaker could come from their gestures as well as facial motion. Here, the data used was more challenging since the speakers did not face the cameras frontally and background noise, such as a ringing mobile phone, was also possible. However, the speakers were required to minimise speaking over each other so maintaining a natural exchange was difficult.

Here we use a much larger audio-visual data set of almost 2 hours with more challenging, complex, and noisy data. The AMI corpus [2] consists of audio-visual data captured from 4 participants in a natural meeting scenario. There are props such as a table, white board and slide screen in the meeting room (see Figure 1), which encourages natural discussions and thus more difficult scenarios. During these natural meetings, speakers could interrupt, talk over each other, and move spontaneously.

We present firstly our experiments on estimating the dominant person from a single distant microphone. Then, we describe two sets of experiments to estimate the audio-visual speaker association. First, personal audio sources and manually annotated ground truth speaker segmentations are correlated with visual activity features from personal close-view cameras. Then, we use a single audio source and investigate how varying the experimental conditions and diarization strategies could affect the output of the audio signal for audio-visual speaker association compared to estimating the dominant person in the meetings. Previously, we found that the dominance estimation was not particularly sensitive to the error rate of the diarization output [6] or a low signal to noise ratio (SNR) from the input source. We examine the extent to which this insensitivity is observed when mapping the speech features to unlabelled video channels and how this affects the dominance estimation task.

The rest of this paper is organised as follows: Section 2 describes the data and dominance annotation procedure; Section 3 describes the different strategies we used for make the speaker diarization algorithm more efficient; Section 4 describes how we extract computationally efficient visual activity features from compressed video taken from individual close-view cameras; Section 5 provides the results from estimating the dominant person with the diarization; Section 6 describes our audio-visual association method; Section 7 provides the results from associating clusters with visual activity features; We conclude in Section 8.

2. Meeting Data and Dominance Annotations

We used the publicly available AMI meeting corpus [2] which was captured in room as shown in Figure 1. A microphone array and four individual close-view cameras were set on the table and each person wore an omni-directional headset and lapel microphone. Non-scripted meeting data was produced where four members of a team were asked to design a remote control device over a series of sessions. This encouraged natural interactions between participants.

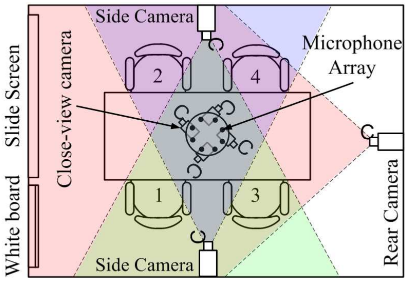


Figure 1. Plan view of the meeting room set up.

59 5-minute non-overlapping meeting segments from 11 sessions were used for human annotations of dominance. 21 annotators were divided into groups of 3 such that each group always annotated the same segments. For each watched segment, annotators were asked to rank the participants according to their level of perceived dominance, from 1 (most) to 4 (least). Annotators were not given a prior definition of dominance. More details about the annotation procedure can be found in [10].

34 meetings had full agreement among the 3 annotators about the most dominant person. For the audio-visual association experiments, we used a subset of these meetings where participants were seated all the time. This consisted of 21 meetings, representing almost 2 hours of data.

3. Speaker Diarization

We use the speaker diarization method in [5] which employs an agglomerative clustering method to merge pairs of speaker clusters iteratively according to the pairwise Bayesian Information Criterion (BIC) score, which is also used to determine when merging should stop. Each cluster is represented as a Gaussian Mixture Model (GMM) of frame-based cepstral features (MFCCs). Calculating the BIC score for potential merge candidates is time-consuming and can be made more efficient by removing unlikely merge hypotheses with a faster scoring method. With these speed-based improvements, speaker diarization outputs were extracted using increasingly faster versions of the algorithm. Robustness testing was also performed by simulating different single distant microphone sources with decreasing signal to noise ratio (SNR). More details about the feature extraction process can be found in [5].

Audio Experimental Conditions : The various experimental conditions that we used can be categorized into a single distant microphone case and a individual close-talk microphone as summarized in Table 1. For the first case, a single audio stream was created by mixing individual close-talk microphone data, i.e. 'Mixed Headset' or 'Mixed Lapel' using a summation. For the latter condition, a single microphone was selected from a microphone array from either the table or ceiling sources.

Mixed Individual Close-talk Microphone	**Single Distant Microphone**
A: Mixed Headset	**C**: Single Table Microphone
B: Mixed Lapel	**D**: Single Ceiling Microphone

Table 1. Summary of various experimental conditions.

As mentioned before, the agglomerative clustering method is data-driven, so it is possible for the algorithm to stop when the number of clusters is not equal to the number of participants. In our previous work [6], we introduced experiments where the final number of clusters was fixed at 4 as well as allowing the algorithm to automatically select the final number of clusters. These experiments did not account for cases when one or more of the participants did not speak in the 5-minute meeting segment. Therefore, we re-adjusted the rule by allowing the algorithm to stop naturally before enforcing further iterations to merge pairs of speaker clusters if the number was greater than 4.

In our previous experiments [6], the diarization results were computed using the original meeting sessions which lasted between 15 and 35 minutes. Speaker segmentations were then generated by dividing the output into 5-minute non-overlapping segments to match those used in the human annotations of dominance. Therefore, information outside of the meeting segment in question was also used in the diarization process. To ensure a fairer experimental set-up, we re-ran the diarization method such that each output was generated solely from the corresponding 5-minute meeting segment. The performance of speaker diarization is measured by the Diarization Error Rate (DER), which is the sum of missing speech, false alarms, and speaker cluster error generated from mapping the clusters to the ground truth speaker segmentations using a dynamic programming method.

A summary of the DER using the shorter 5-minute meeting segments is shown in Table 2 which shows the different experimental conditions and their corresponding DERs, signal-to-noise ratios (SNRs) and speed increases relative to real-time. The terms 'KLFM', 'PCFM', and 'NoFM' refer to the KL-divergence Fast Matching, Pitch Correlogram Fast Matching, and No Fast Matching respectively (see [5]). Conceptually these computationally efficient strategies can be thought of as fast, medium, and slow methods. In an experiment performed on the mixed lapel data, the 'NoFM' baseline system ran at $1.22\times$ real-time (RT), improved to $1.0\times$RT for the PCFM strategy and further improved to $0.77\times$RT for the KLFM case using the same experimental set up as we had used previously [5]. The rows and columns of Table 2 have been labeled with letters and numbers for easy reference. We can observe a decrease in the average performance as well as a reduced sensitivity to the SNR of the input signal compared to the chunked session-based DERs in our previous experiments [6].

Source	SNR (dB)	**Number of speaker clusters $\leq$ 4**			**Automatic speaker cluster estimation**		
		KLFM	PCFM	NoFM	KLFM	PCFM	NoFM
A	31	33.17	32.17	32.52	33.78	32.83	33.16
B	22	34.71	34.19	34.94	36.47	35.91	36.35
C	21	35.34	34.94	34.94	36.14	36.19	36.16
D	18	35.94	36.22	34.85	35.96	36.89	36.55
		1	**2**	**3**	**4**	**5**	**6**

Table 2. Diarization results (DER) where the labels **A**-**D** refer to the experimental conditions described in Table 1.

4. Computationally Efficient Video Features

To estimate the audio-visual association, we used a frame-based visual activity feature that can be matched with the speaking activity patterns. Visual activity features are extracted from personal close-view cameras by re-using some of the video processing used for video compression [3]. We extracted the motion vector magnitude and the residual coding bit-rate to construct an estimate of personal activity levels using the method detailed in [7]. These features are illustrated in Figures 2 (b) and 2 (c) respectively.

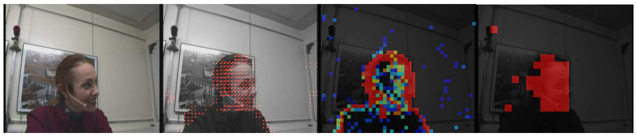

(a) (b) (c) (d)

Figure 2. Compressed domain video feature extraction. (a) Original image, (b) Motion vectors, (c) Residual coding bit-rate, (d) skin-coloured regions.

For each camera view, we estimate a participant's activity levels by implementing a block-level skin-colour detector working mostly in the compressed domain which can detect head and hand regions as illustrated in Figure 2 (d). To do this, we use a GMM to model the distribution of chrominance coefficients [12] in the YUV colour-space. Specifically, we model the chrominance coefficients, (U, V), as a GMM, where each Gaussian component is assumed to have a diagonal covariance matrix. In the Intra-frames, we compute the likelihood of observed chrominance DCT DC coefficients according to the GMM and threshold it to determine skin-colour blocks. These blocks in the Inter-frames are inferred by using motion vector information to propagate them through the duration of the group-of-picture (GOP). In the presence of long GOPs, such as in the AMI meeting videos, accumulated errors could lead to large areas of the frame being falsely detected as skin-colour blocks. To prevent this, we add an additional verification step, performed in the pixel domain, to remove blocks that are erroneously tagged as skin-colour blocks; this verification step is performed only if a block is suspected to be a skin block.

For each frame the average motion vector magnitude or residual coding bit-rate over all the estimated skin blocks is calculated and used as a measure of individual visual activity. The novelty of using these as visual activity features for speaker association is that they are block-based and are already computed during video compression. Compared to extracting many higher resolution pixel-based features such as optical flow, compressed-domain features are much faster to extract, with a run-time reduction of 95%.

While personal close-view cameras are used, the distance from the camera causes scale and pose issues, as shown in some example shots in Figure 3. By averaging activity measures over detected skin-colour blocks, we hope to mitigate some of these issues. In addition, we also know that people move even when they are not speaking, which makes associating these visual activity features with the estimated speaker clusters challenging. The video capture results in 4 streams; one for each close-view camera. Each of the 4 streams were represented either using the motion vector magnitude features (Vector) or residual coding bitrate (Residue) with filtering of the skin-coloured regions.

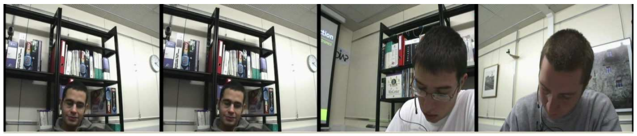


Figure 3. Example screen-shots from the close-view cameras.

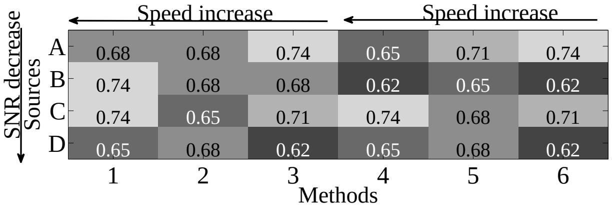
5. Estimating the Dominant Person

The results for estimating the most dominant person are summarised in Table 3 where the best and worst results were 74% and 62% respectively. Interestingly, in three out of the four cases where 74% of the estimates were correct, the fastest diarization strategy was used while three out of the four worst results were produced from the slowest method. Forcing the final number of clusters to be less than or equal to 4 did not seem to affect the results. The results showed a lack of sensitivity to the SNR compared to the DERs shown in Table 2. This meant the highest performance was also produced using an audio signal with the second worst SNR and the fastest diarization strategy. Compared to the baseline result of 85% which was estimated using the headset speaker segmentations, there was a decrease in performance when noisy speaker segmentations generated from the various speaker diarization methods were used.

6. Associating Speaker Clusters with Unlabelled Video Channels

6.1. A Naïve Case

We used ground truth speaker segmentations (GT) created from manual annotations for our initial association analysis. In addition, to compare with the baseline results presented for the most dominant person estimation task, speaker segmentations from the audio signal taken from personal headset microphones (HS) were also generated. These were associated with the two real-valued visual activity features using the residual coding bit-rate (Residue) or motion vector magnitudes (Vector). The headset segmentations were generated by extracting the speaker energy from each headset and then thresholding this value to create a binary signal where 1 represents speaking and 0 is silence.



	1	2	3	4	5	6
A	0.68	0.68	0.74	0.65	0.71	0.74
B	0.74	0.68	0.68	0.62	0.65	0.62
C	0.74	0.65	0.71	0.74	0.68	0.71
D	0.65	0.68	0.62	0.65	0.68	0.62

Table 3. Colour-coded representation of the results for the dominant person task using diarization clusters generated from 5-minute meeting segments where higher performance is shaded lighter. The numbers are shaded differently for clarity only.

For each pair-wise combination of speaking and visual activity channels, their corresponding normalised correlation was calculated. We then matched the channels by using an ordered one-to-one mapping based on associating the best correlated channels first. Three different evaluation criteria were used to observe the differences in discriminability of the data by varying the leniency of the scoring into soft, medium and hard criteria : EvS gives each meeting a score of 1 if at least 1 of the 4 speech and visual activity channels match correctly; EvM scores 1 if at least two of the channels match correctly; EvH scores 1 only if all 4 visual activity channels are assigned correctly. The proportion of correctly associated meetings using both visual activity feature types are shown in Table 4 below. Surprisingly, correlating the headset segmentations and Residue visual activity channels performed best, though the difference in performance differs at most, by 2 meetings. Also, it was also encouraging to see that even for the hard evaluation strategy, the performance remained high for this case.

Audio	**Visual**	EvS	EvM	EvH
GT	**Residue**	1	1	0.86
	Vector	0.95	0.95	0.71
HS	**Residue**	1	1	0.9
	Vector	1	0.95	0.81

Table 4. Proportion of correctly associated speech segmentations generated from associating visual activity from the close-view cameras using (i) the ground truth speaker segmentations and also (ii) individual headset microphones. See text for descriptions of the evaluation criteria.

6.2. Evaluating Speaker-Cluster to Video Mappings

The results shown in the previous subsection used relatively clean segmentations generated from individual headset segmentations or from the ground truth. When the clusters from the speaker diarization output is used, we can expect two issues to arise. Firstly, the number of speaker clusters from the speaker diarization engine could be unequal to the true number of participants. Secondly, we must quantify the quality of the mappings. To do this, we computed the pair-wise normalised correlation between (i) the speaker clusters and visual activity features and (ii) the speaker clusters and either the ground truth speaker segmentations or those extracted from the headset microphones. The mappings for both cases were calculated again based on an ordered one-to-one mapping starting from the pair with the highest correlation. If there were fewer speaker clusters than motion channels, mappings were forced to ensure each motion channel mapped to a speaker cluster.

In reality, there can also be more clusters than participants so more than one speaker cluster can be associated with the same motion channel. We accounted for this by running the association algorithm over all possible pair-wise cluster-video combinations as a first-pass. Once 4 mappings were found, these were set aside and the remaining clusters were mapped afresh to all 4 possible visual activity channels using the same strategy, where the cluster-visual activity channels with the highest normalised correlation were matched first using the same ordered one-to-one mapping procedure. After the second pass, a visual activity channel can be mapped to one speaker cluster in the first and another in the second, facilitating many-to-one mappings.

Using the mapping of the clusters to labelled speaker segmentations, a scoring criteria is enabled where the mapping is true only when the corresponding GT or HS segmentation is associated with the correct visual activity channel through the corresponding speaker cluster. We used again the three evaluation criteria EvH, EvM, and EvS, which assigns respectively a score for each meeting only when all, at least two, or at least one of the mappings is correct. A fourth evaluation criterion, $EvFu$ was introduced to account for fuzzy cases where more than one GT or HS speaker channel is associated to the same video channel. Each meeting can have a maximum score of 4; each correct match, at most 1. For each correct mapping, the score is calculated as the reciprocal of the total number of participants that have been associated with that particular visual activity channel. Finally, we integrated these association results back into the dominance task by checking all correct mappings to see if they matched with the longest cluster length (which we expect to be the most dominant person).

7. Results

The speech-visual activity association was performed on 21 5-minute segments where all the participants were always seated in their close-view camera. We tested using the speech activity output generated from all the different speaker diarization strategies and conditions described in Section 3. Finally, we used both the ground truth segmentations and headset segmentations to evaluate the mappings. In all, 16 different combinations of evaluation criteria and reference segmentations were used and for each of these combinations, we had 24 different experimental conditions for the diarization output. From Table 4, we decided only to use the Residue version of the visual activity features since they gave the best performance. Figure 4 shows a summary of all the speech-motion association results using the ground truth (GT) or headset (HS) segmentations and the 4 evaluation criteria EvH, EvM, EvS and EvFu.

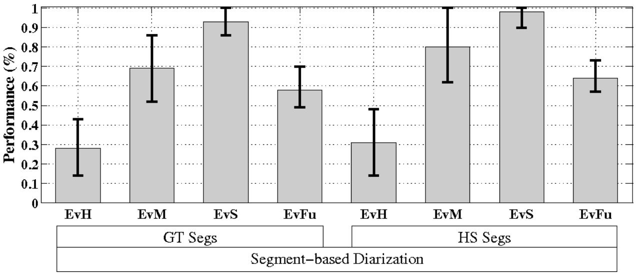


Figure 4. Average, lowest, and highest performance results for each experimental condition and evaluation strategy.

Table 5 shows the results using our 4 evaluation strategies and different reference speaker segmentations. There is a slight improvement in performance when using the headset rather than ground truth speaker segmentations to evaluate the quality of the clusters. Also, a degradation in performance is observed as the evaluation criteria becomes more strict, as highlighted in Figure 4. The best average score was achieved by the EvS HS case with an average and highest performance of 93% and 100% respectively. Closer inspection of the experiments showed that there were 5 cases where 100% performance was achieved using EvS HS (and 2 for the EvS GT case). These were mostly clusters created using the diarization strategy with a medium speed increase (PCFM). In 4 of the cases, the input audio source used a true distance microphone source and 3 estimated the number of speaker clusters automatically. For the softer scoring criterion, $EvFu$, we found that the performance was better than the corresponding results using EvH, indicating that there were a number of cases where there were many-to-one or many-to-many mappings.

	GT Speaker Segs				**HS Speaker Segs**			
	EvH	EvM	EvS	$EvFu$	EvH	EvM	EvS	$EvFu$
Average	0.28	0.69	0.93	0.58	0.31	0.8	0.98	0.64
Max	0.43	0.86	1	0.7	0.48	1	1	0.73
Min	0.14	0.52	0.86	0.49	0.14	0.62	0.9	0.57

Table 5. Summary of the performance using all 4 performance evaluation strategies and either the ground truth or automatically generated headset speaker segmentations. The evaluation criteria are as described before.

Finally, Table 6 shows the percentage of meetings where the association of the longest speaker cluster with the correct visual activity channel was made. Again, experiments were conducted with either ground truth and headset speaker segmentations for evaluation. We were unable to complete a thorough examination of these experiments since the data used for the dominance estimation task and cluster-video association did not overlap fully. 15 out of the 34 meetings where the most dominant person was selected by all 3 annotators also contained seated participants for the entire segment. To maximise the number of samples, we used the same subset of 21 meetings as those used for the audio-visual association experiments. The best performing dominance and association results were achieved by using the headset segmentations with an average performance of 70% where there were 3 cases where a performance of 86% was achieved. Two of these cases used the true single distant microphone sources (C and D) and one of these used the fastest diarization strategy. However, in each of these three best cases, the number of speaker clusters was fixed to be less than or equal to four so it was necessary to know the true number of participants a priori. Again, using the ground truth segmentations lead to slightly worse results. On closer inspection there were some experimental conditions where using the headset speaker segmentations led to the correct pairing of speaker clusters and motion channels in 6 more meetings compared to the same conditions using the ground truth speaker segmentations. It is interesting to note also that when the same experiments were run using the diarization output from chunked version of the longer meeting sessions, there was one case

	GT	HS
Average	0.64	0.7
Max	0.81	0.86
Min	0.52	0.52

Table 6. Percentage of meetings where the correct mapping was given to the cluster with the longest speaking length.

8. Conclusion

We conducted experiments to investigate the challenges in identifying the most dominant person in meetings using both audio and video data. To our knowledge, our experiments on audio-visual association on almost 2 hours of data, used the highest degree of complexity in terms of the number of people that needed to be matched to a single audio source. Our results show that it is possible to associate the dominant speaker with a set of visual activity candidates quite robustly using a simple greedy mapping method. Although there was a decrease in performance in both the DER, dominance performance and speaker-visual activity mappings when the diarization output was computed from shorter meeting segments, this decrease was not always significant and showed some surprisingly good results even with an input audio signal with the worse SNR. In terms of computational efficiency, and practical use, conditions (D,1) provided the best all-round performance for estimating the dominant person audio-visually. We acknowledge that out experiments had few data points so it is difficult to draw strong conclusive remarks about the differences in performance between the strategies. To assess the performance of both the dominance and audio-visual association more fully, we will need to represent each person's motion effectively when they are not seated so that a larger data set can be tested.

Acknowledgments This research was partly funded by the US VACE programme, the EU project AMIDA (pub. AMIDA-103), the Swiss NCCR IM2, and Singapore A*STAR. We also thank Gerald Friedland (ICSI) for providing some code.

References

[1] S. Basu, T. Choudhury, B. Clarkson, and A. Pentland. Learning human interactions with the influence model. In *NIPS*, 2001.

[2] J. Carletta, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos, W. Kraiij, M. Kronenthal, G. Lathoud, M. Lincoln, A. Lisowska, M. McCowan, W. Post, D. Reidsma, and P. Wellner. The ami meeting corpus: A pre-announcement. In *Proc. MLMI*, 2005.

[3] S.-F. Chang. Compressed-domain techniques for image/video indexing and manipulation. In *Proc. IEEE ICIP*, pages 314–317, 1995.

[4] N. E. Dunbar and J. K. Burgoon. Perceptions of power and interactional dominance in interpersonal relationships. *Journal of Social and Personal Relationships*, 22(2):207–233, 2005.

[5] Y. Huang, O. Vinyals, G. Friedland, C. Müller, N. Mirghafori, and C. Wooters. A fast-match approach for robust, faster than real-time speaker diarization. In *ASRU*, 2007.

[6] H. Hung, Y. Huang, G. Friedland, and D. Gatica-Perez. Estimating the dominant person in multi-party conversations using speaker diarization strategies. In *ICASSP*, 2008.

[7] H. Hung, D. Jayagopi, C. Yeo, G. Friedland, S. Ba, J.-M. Odobez, K. Ramchandran, N. Mirghafori, and D. Gatica-Perez. Using audio and video features to classify the most dominant person in a group meeting. In *ACM Multimedia*, 2007.

[8] J. W. F. III and T. Darrell. Speaker association with signal-level audiovisual fusion. *IEEE Transactions on Multimedia*, 6(3):406–413, 2004.

[9] J. W. F. III, T. Darrell, W. T. Freeman, and P. A. Viola. Learning joint statistical models for audio-visual fusion and segregation. In *NIPS*, pages 772–778, 2000.

[10] D. B. Jayagopi, H. Hung, C. Yeo, and D. Gatica-Perez. Modeling dominance in group conversations using non-verbal activity cues. IDIAP Research Report, December 2007.

[11] M. S. Mast. Dominance as expressed and inferred through speaking time. *Human Communication Research*, (3):420–450, July 2002.

[12] S. J. McKenna, S. Gong, and Y. Raja. Modelling facial colour and identity with gaussian mixtures. *Pattern Recognition*, 31(12):1883–1892, 1998.

[13] K. Otsuka, J. Yamato, Y. Takemae, and H. Murase. Quantifying interpersonal influence in face-to-face conversations based on visual attention patterns. In *Proc. ACM CHI Extended Abstract*, Montreal, Apr. 2006.

[14] R. Rienks, D. Zhang, D. Gatica-Perez, and W. Post. Detection and application of influence rankings in small group meetings. In *ICMI '06: Proceedings of the 8th international conference on Multimodal interfaces*, pages 257–264. ACM Press, 2006.

[15] E. Rosa and A. Mazur. Incipient status in small groups. *Social Forces*, 58(1):18–37, September 1979.

[16] M. Siracusa and J. Fisher. Dynamic dependency tests for audio-visual speaker association. In *ICASSP*, April 2007.

[17] H. Vajaria, T. Islam, S. Sarkar, R. Sankar, and R. Kasturi. Audio segmentation and speaker localization in meeting videos. In *ICPR*, pages 1150–1153, 2006.

[18] D. Zhang, D. Gatica-Perez, S. Bengio, and D. Roy. Learning influence among interacting Markov chains. In *NIPS*, 2005.