Chapter 1
Computationally Efficient Clustering of Audio-Visual Meeting Data

Hayley Hung[1], Gerald Friedland[2], and Chuohao Yeo[3]

Abstract This chapter presents novel computationally efficient algorithms to extract semantically meaningful acoustic and visual events related to each of the participants in a group discussion using the example of business meeting recordings. The recording set-up involves relatively few audio-visual sensors, comprising a limited number of cameras and microphones. We first demonstrate computationally efficient algorithms that can identify who spoke and when, a problem in speech processing known as speaker diarization. We also extract visual activity features efficiently from MPEG4 video by taking advantage of the processing that was already done for video compression. Then, we present a method of associating the audio-visual data together so that the content of each participant can be managed individually. The methods presented in this article can be used as a principal component that enables many higher-level semantic analysis tasks needed in search, retrieval, and navigation.

With the decreasing cost of audio-visual sensors and the development of many video-conferencing systems, a growing trend for creating instrumented meeting rooms could be observed. As well as aiding teleconferencing, such meeting rooms could be used to record all meetings as a tool for staff training and development or to remind them of certain agenda items that were discussed. Given the number of meetings that occur for a single person or even a work group, recording and storing meetings alone would not be useful unless they could be searched and browsed easily later.

In this chapter, we discuss ways in which we can move towards the use of instrumented meeting rooms while also minimizing the amount of audio-visual sensors, thus enabling fast set-up and portability; We show experiments to cluster the audio and visual data of each person where only one microphone and two cameras are used to record the group meetings. From this, we present computationally-efficient algorithms for extracting low-level audio and video features. The chapter is divided

[1]University of Amsterdam, Netherlands · [2]International Computer Science Institute (ICSI), Berkeley, USA · [3] Institute for Infocomm Research, Singapore

into sections describing firstly the general challenges of the meeting room scenario and what types of applications have been proposed. Then, we describe the related work on audio-visual speaker segmentation and localization in Section 1.1. In Section 1.2, we describe the overall approach that is presented in this chapter. Then, we describe the audio-visual sensor set-up that we used in evaluating our algorithms in Section 1.3. Next, we describe how speakers and their turn-taking patterns are extracted using an on-line speaker diarization algorithm (Section 1.4). Then, in Section 1.5 we describe how visual activity from individuals can be extracted from compressed-domain features and compare this to conventional pixel- domain processing. In Section 1.6, we describe a method of associating audio-visual data and present bench-marking results. We conclude in Section 1.8 and discuss the future challenges.

1.1 Background

Clustering audio-visual meeting data can involve the grouping of events on different levels. From the coarsest level, we may want to group them based on date, location or which work-group participated. If we increase the granularity, we observe events within a single meeting such as the types of activities that took place. Increasing the granularity further, each activity consists of a conversation type (ranging from monologue to discussion) where speech turn-taking events occurs. For each speech event, there are also accompanying motion features such as a nod of the head, that might accompany a statement of agreement. We can go further in granularity by observing each speech utterance such as separation into phonemes. The motion can be organized based on the types of motion that occur such as whether it is an upward or downward motion.

Like any data mining task, our ultimate obstacle in creating a system that can cater completely to our searching and browsing needs is the problem of the *Semantic Gap*. The semantic gap is defined as the difference between the cognitive representation of some data compared to what can be extracted in terms of its digital representation. In this chapter, we concentrate on discussing how audio-visual meeting data can be clustered by who spoke when and where. The approach we present here, consists of two tasks. The first clusters audio data based on how many speakers there are and when they speak. Semantically, this is not so meaningful since we only know that there are N speakers and when they spoke but we don't know who each speaker was. The second task takes these speaker clusters and identifies where they are in a set of video streams by associating the clustered audio with video features, which can then be used to show the corresponding speakers at the relevant time. This step already closes the semantic gap in terms of finding speakers and when they spoke and provides audio and video footage of how a speaker delivered a line.

Historically, speaker diarization has been a useful tool for the speech processing community since once the speakers have been identified, automatic speech recognition (ASR) can be applied to the utterances and attributed to a particular person. There are many who believe that closing the semantic gap has involved processing speech in terms of its verbal content. From a linguistic viewpoint, this seems to be

the natural choice if we wish to extract the verbal content of what is being said so that interactions can be analyzed semantically. However, while semantics are closely related to verbal cues, meaning can also be extracted from non-verbal features. In some cases, the non-verbal cues can be a better indicator of the sentiment of the delivery of a phrase. A common example would be the use of sarcasm where someone may say "yes" when they actually mean "no". Analyzing the verbal content alone would provide us with the incorrect interpretation of the message but looking at the non-verbal cues, we might see that the delivery contained audio features that are more highly correlated with disagreement.

Practically speaking, systems that can automatically analyze audio-visual data using ASR and computational linguistics face many challenges. In natural speech, people do not always speak in perfect sentences and may correct themselves, change topic, talk over each other or complete each other's sentences. Typically ASR algorithms are plagued with challenges such as variations in accent, overlapping speech, and differences in delivery of the same word from the same person (which can depend on the preceding and following words), errors from detected words which are out of vocabulary, or inaccurate language models. The state-of-the art word error rate (WER) using distant microphones is around 25% using close-talk head-set microphones and around 40% using a distant (0.5m) microphone source [29]. In terms of computational linguistics, analyzing dialog acts (the aim of the utterance e.g. agreement, disagreement, knowledge transfer), summarization, topic detection or the sentiment of what was said based on the ASR output can introduce further errors into the system chain. This is particularly problematic if the content of the exchanges are to be used for the analysis of higher semantic concepts from the data.

Analyzing or identifying these higher semantic concepts goes beyond the traditional meeting browsing technologies that can be used to navigate between changes in topic in a conversation or simple functions just as skipping through a video every 5 minutes. Being able to analyze a meeting by its social non-verbal content takes the potential of meeting browsing technology to a more intuitive level for users. Much of data mining and audio-visual clustering has been treated as a data-driven problem but perhaps in the context of recorded meetings and in particular where conversations are concerned, we must not overlook the stable nature of the non-verbal behavior that is exhibited during these interactions. For example, it is known that we move more than our mouths when we talk; we gesticulate for emphasis or to help us get our point across[43]. If our final goal is to browse meeting data in terms of social memory triggers, can the patterns of non-verbal behavior seen in social interactions be used to cluster the data too? That is, could aspects of non-verbal behavior during conversations provide us a simple and practical solution to this problem? Recent work on estimating behavioral constructs such as find who is dominant [35], the personality of participants [52] or what roles people have [20] suggest that using automatically extracted non-verbal cues can be effective.

For meetings in a natural setting, we expect to see mostly unconstrained conversations. Natural conversation in meetings involve many factors that are generally unwanted in a clean test scenario. The first is overlaps or interruptions in speech. Traditional data-sets [50] that are used to test audio-visual synchrony algorithms

assume that only one person speaks at a time. In more complex cases, one person mouths words not corresponding to the associated audio sequence in order to confound simpler synchrony algorithms. Others contain subjects reciting digits simultaneously. However, in all cases, the speech is not natural and test data in such conditions do not reflect the majority of circumstances in which people find themselves talking.

Other aspects of real conversations involves natural body movements. In natural conversations, people move to aid emphasis of what they are saying, provide feedback for others and regulate their gaze patterns to encourage a smooth flow of conversation between conversants [43, 30]. Promising work has been presented to take advantage of the correlation between more holistic body motion and speech [59, 60, 32, 31]. Such methods have shown a relationship between global body motion and speech over longer term sequences. The experiments presented in this chapter, continues in this direction, exploring the extent to which we can use findings in the psychology literature to address the audio-visual clustering problem in meetings more directly for constructing a plausible practical approach to the problem of speaker localization. For the remainder of this section, we will discuss firstly the general challenges faced with organizing meeting data. Then we will concentrate the discussion on related work on speaker diarization and on audio-visual synchrony, related to speech and finally some background on the findings in psychology on audio-visual synchrony during conversations.

1.1.1 Challenges in Meeting Analysis

Organizing audio-visual meeting data involves using many different sorting criteria. For now, let us concentrate on scenarios where all the conversants are co-located so that interactions can occur face-to-face. Even under such circumstances where acoustic and lighting conditions can be controlled, there are still considerable challenges that can be addressed in a multi-disciplinary domain from signal processing, to computer vision, linguistics, and human-computer interaction.

Activities in meetings consist mainly of conversations or interactions between the participants. Within meetings, people can communicate with each other in different permutations and at different times. They can talk over each other, have subconversations, be involved in multiple conversations at the same time, and can provide verbal as well as non-verbal signals to others. In some cases the verbal and non-verbal delivery of a message can be contradictory.

As well as investigating group conversational dynamics in the work place from a psychological perspective [6, 49, 17], work has been done in the domain of computational modeling [40, 48, 2, 54]. Due to European project initiatives, the computational modeling of meetings has been considered in terms of either visual or audio-visual segmentation of the group activities as discussions, monologues, notetaking, presentations or writing on a white board from the Multi-Modal Meeting Manager Corpus (M4) (http://www.idiap.ch/mmm/corpora/m4-corpus/) [40, 2, 54] where the meetings were scripted so each meeting activity and the times of execution were predetermined. The main problem with approaching meeting analysis

from this perspective is that in reality, it is very difficult to objectively label monologues, dialogues, discussions, or presentations. For example, if someone is giving a presentation and someone else asks a question, which ultimately leads to a discussion, then is the current scenario a presentation or a discussion? The answer lies in the interval over which the judgment is made or the temporal context which is applied. Therefore, depending on whether the judgment is made on a fine-grained time scale or a longer time scale, the judgment of the scenario can also be different. Since the M4 corpus, new audio-visual meeting data (Augmented MultiParty Interaction (AMI) corpus http://www.idiap.ch/mmm/corpora/ami) has been recorded, where the scripting part of the scenario was removed. In more natural meeting scenarios, people do not cut from doing a presentation to a discussion or a monologue necessarily so annotating these meetings in terms of meeting actions is not practical.

With this in mind, it is probably easier to extract semantically meaningful features which are easier to evaluate. The problem with analyzing meeting actions is that labeling is strongly dependent on the temporal context. Rather than examining temporal intervals of time, we can also segment based on events such as a change of speaker or when someone starts or stops speaking. Such instantaneous events are much less ambiguous to label. This can be done by either speech/non-speech detection for cases where each person has their own microphone [66] or using speaker diarization if a single microphone cannot be directly associated with a single speaker.

If we are able to cluster the audio and video information of a speaker, we can begin to analyze more complex behaviors such as who responds to whom. Analysis of turn-taking patterns in discussions can be quite powerful for indicating who is dominant [35] or what roles people play in a meeting [20, 34]. With an audio-visual clustering method we could automatically obtain both the audio and video information for the project manager for a meeting, for example. Given that the discussion above has established that it is easier to analyze meetings in terms of these turn-taking events, we provide a background review of speaker diarization. In addition, we provide a review of work on the audio-visual association of speakers so that some semantic meaning can be associated with the speakers that are identified. Finally, we provide some background information about how human body motions are related to speech during conversations.

1.1.2 Background on Speaker Diarization

The goal of speaker diarization is to segment audio into speaker-homogeneous regions with the ultimate goal of answering the question “who spoke when?” [55]. While for the related task of speaker recognition, models are trained for a specific set of target speakers which are applied to an unknown test speaker for acceptance (the target and test speaker match) or rejection (mismatch), in speaker diarization there is no prior information about the identity or number of the speakers in the recording. Conceptually, a speaker diarization system therefore performs three tasks: First, discriminate between speech and non-speech regions (speech activity detection); second, detect speaker changes to segment the audio data; third, group the segmented regions together into speaker-homogeneous clusters.

Some systems combine the two last steps into a single one, i.e. segmentation and clustering is performed in one step. In the speech community, different speaker diarization approaches have been developed over the years. They can be organized into either one-stage or two-stage algorithms, metric-based, and probabilistic systems, and either model-based or non-model-based systems.

Many state-of-the-art speaker diarization systems, use a one-stage approach, i.e. the combination of agglomerative clustering with Bayesian Information Criterion (BIC) [12] and Gaussian Mixture Models (GMMs) of frame-based cepstral features (MFCCs) [55] (see Section 1.4). Recently, a new speaker clustering approach, which applies the Ng-Jorden-Weiss (NJW) spectral clustering algorithm to speaker diarization is reported [45].

In two-stage speaker diarization approaches, the first step (speaker segmentation) aims to detect speaker change points and is essentially a two-way classification/decision problem, i.e., at each point, a decision on whether it is a speaker change point or not needs to be made. After the speaker change detection, the speech segments, each of which contains only one speaker, are then clustered using either top-down or bottom-up clustering.

In model-based approaches, pre-trained speech and silence models are used for segmentation. The decision about speaker change is made based on frame assignment, i.e. the detected silence gaps are considered to be the speaker change points. Metric-based approaches are more often used for speaker segmentation. Usually, a metric between probabilistic models of two contiguous speech segments, such as GMMs, is defined and the decision is made via a simple thresholding procedure.

Over the years, research has concentrated on finding metrics for speaker change detection. Examples are the Bayesian Information Criterion (BIC) [12], cross BIC (XBIC) [36][4], Generalised Likelihood Ratio (GLR) [18], Gish distance [26], Kullback-Leibler distance (KL) [9], Divergence Shape Distance (DSD) [39]. A more detailed overview can be found in [3]. Newer trends include the investigation of new features for speaker diarization, such as [24], [61], and novel initialization methods.

1.1.3 Background on Audio-Visual Synchrony

So far, the speaker diarization system provides some intervals of speech associated with a single person but we do not have information about what they look like or how the message was delivered non-verbally. This can be done by associating the audio streams with the correct video stream by identifying or exploiting the synchrony between the two modalities. Alternatively, sound source localization from video can be used to tackle a similar problem. Most computational modeling has involved identifying one or two people in a single video camera only where short term synchrony of lip motion and speech are the basis for audio-visual localization. Audio-visual synchrony or sound source localization can be considered a task in itself. However, both these tasks could be combined and recent work has started to consider both speaker diarization and localization as a single audio-visual task.

Common approaches to audio-visual speaker identification involve identifying lip motion from frontal faces [46], [47], [13], [22], [21], [53], [57], [58]. Therefore, the underlying assumption is that motion from a speaker comes predominantly from the motion of the lower half of their face. This is further enforced by artificial audio-visual data of short duration, where only one person speaks. In these scenarios, natural conversation is not possible so problems with overlapping speech are not considered. In addition, gestural or other non-verbal behaviors associated with natural body motion during conversations are artificially suppressed [50].

Nock et al.[46] presents an empirical study to review definitions of audio-visual synchrony and examine their empirical behavior. The results provide justifications for the application of audio-visual synchrony techniques to the problem of active speaker localization in the more natural scenario of broadcast video. Zhang et al. [69] presented a multi-modal speaker localization method using a specialized satellite microphone and omni-directional camera. Though the results seem comparable to the state-of-the-art, the solution requires specialized hardware, which is not desirable in practice. Noulas et al. [47] integrated audio-visual features for on-line audio-visual speaker diarization using a dynamic Bayesian network (DBN) but tests were limited to two-person camera views. Tamura et. al. [58] demonstrate that the different shapes the mouth can take when speaking facilitates word recognition under tightly constrained test conditions (e.g., frontal position of the subject with respect to the camera while reading digits).

The approaches discussed above were often tested on very limited data sets (which are not always publicly available) and were often recorded in highly constrained scenarios where individuals were unable to move or talk naturally. In general, the speakers face the camera frontally and do not talk over or interrupt each other. In contrast to previous methods which combine audio and video sources in the early stages of the speaker diarization process, we present a late fusion approach where noisy video streams are associated with estimated speaker channels.

In terms of finding speakers in conversational settings where video data does not capture high-resolution faces, Vajaria et al. [59, 60] were the first to consider the global body motion could be synchronous with speech. They presented a system that combines audio and video on a feature-level using eigen-vector decomposition of global body motion. Hung et al. [31] developed this notion further by considering how simple motion features could be used to identify speakers in video streams for group discussions. Finally Campbell and Suzuki [10] analyzed speech and upper torso motion behavior in meetings to study participation levels but did not go further into evaluating how well speech and motion could be correlated.

1.1.4 Human Body Motions in Conversations

In contrast to much previous work in this area, we have found that relying on lip motion to identify speakers is not always necessary and is not always possible [32, 31]. In the psychology literature, it has been shown on many occasions that speaker and also listener movements are directly related to the role they play in a conversation

[37, 43]. We will explore this in more detail here to show that such non-verbal cues play a huge role in understanding and inferring behavior types in conversations.

In social psychology, human body movements in conversations have been studied from different perspectives. The first looks at the movements of speakers, the second, at the movement of listeners, and the final considers the synchrony between the movements of speakers and listeners. The first two are important for understanding what differentiates speakers from listeners in terms of kinesic behavior while the third is used more to measure the degree of mutual engagement between conversants. The latter is beyond the scope of this paper but more details can be found in a critique of work on interactional synchrony by Gatewood and Rosenwein [25].

The first aspect involving the movement of speakers suggests that speakers accompany their speech with gestures [43, 37]. Gestures accompanying speech themselves have been classified in many different ways. Adam Kendon defined gesture as a

> "range of visible bodily actions that are . . . generally regarded as part of a person's willing expression." (p 49).

The reason for gesturing has been explained as a means of increasing precision [43, 27], an evolutionary origin of language [38], or as an aid to speaking to facilitate lexical retrieval [42, 43]. Whatever the reason for moving when speaking, psychologists are in agreement that we definitely move a number of body parts when we speak. Moreover, it was noted by Gatewood and Rosenwein, that "normal human beings exhibit remarkable integration of speech and body motion at the subsecond time scale." (p13, [25]). Such a phenomenon was labeled as 'self synchrony' by Condon and Ogston [15] who later elaborated that,

> "As a normal person speaks, his body 'dances' in precise and ordered cadence with the speech as it is articulated. The body moves in patterns of change which are directly proportional to the articulated pattern of the speech stream....There are no sharp boundary points but on-going, ordered variations of change in the body which are isomorphic with the ordered variations of speech" (p153) [16].

Gestures that accompany speech can be divided into a number of different categories involving manipulation of facial features, head pose, the trunk (or upper torso), arms, shoulders and hands. Hadar et al. found that short and rapid head movements can accompany points of stress in a sentence as a person speaks [27]. In addition, Hadar et al. also found that the frequency of large linear movements of the head was correlated with a person's speaking time in a conversation [28]. In larger groups, speakers can also move their head to address all the participants. Depending on the person's status within the group, their level of conversant monitoring can vary [19].

Hand motions have been shown to be very related to the content of what is being said; it has been suggested by Armstrong et al. that,

> "Most gestures are one to a clause, but when there are successive features within a clause, each corresponds to an idea unit in and of itself. . . Each gesture is created at the moment of speaking and highlights what is relevant. . . " (p 40-41) [5].

McNeill called such gestures 'spontaneous' where "their meaning is determined on-line with each performance." (p 67) [43] and identified four types of relationships between spontaneous gestures and speech; iconic, metaphoric, beat, and abstract deictic. Iconic gestures represent objects and events in terms of resemblance; metaphoric gestures represent an abstraction; beat features are rhythmic movements of the hand such as for counting or indexing a list; and abstract deictics represent locations of objects within a gesture space [43].

The listener in a conversation can, provide feedback to the speaker, indicate that they wish to claim the floor, or indicate their interest in a conversation. It was found by Hadar et al. [27] that listener's head movements tended to involve more "linear and expansive" movements when indicating that they wanted to speak, "symmetric and cyclic" when providing simple feedback such as 'yes' or 'no' responses, and "linear but with shorter movements" during pauses in the other's speech, which could be attributed to 'synchrony' behavior between conversants. While speaker's movements tend to be more pronounced, the movements of listeners are less pronounced but still observable. Harrigan found that body movements occurred more frequently when a person was requesting a turn than during the middle of someone else's speaking turn [30], showing that listeners tend to move less. She also found that hand gestures tended to precede a turn compared to feedback responses that were observed from motion from the head such as nods, shakes and tilts, facial expressions and shoulder shrugs. In particular, gestures from the hands were related to speech, serving to accent or emphasize what was being said.

1.2 Approach

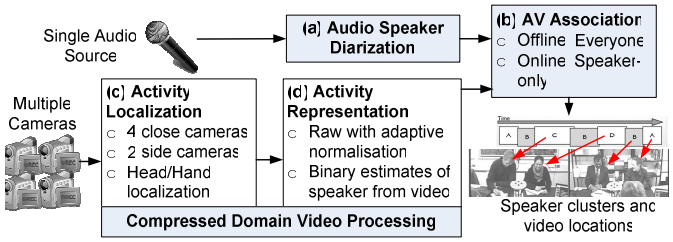


Fig. 1.1 Figure showing our approach. The work consists of two stages: (a) solving the task of 'who is speaking now?' based on audio information only; (b) associating speakers with video streams. Different types of video features (c-d) are used to enhance the practicality, and performance of the system.

Figure 1.1 shows a flow diagram of the approach that we have taken for clustering the audio-visual meeting data in terms of who spoke when and where they are. The goal of the presented system is to identify speakers and their approximate locations in multiple camera streams, in an on-line and real-time fashion. We perform experiments with four-participant meetings for cases where there are either four cameras (one for each person), or two cameras (two people are shown per camera). A summary of the approach is listed below.

(a) On-line real-time speaker diarization: Speaker clusters are generated using the audio data to represent each speaker and when they speak. From this unsupervised data-driven method, a set of speaker clusters are generated where it is assumed that one speaker corresponds to one cluster.
(b) Audio-visual association of speakers streams and video: Using these speaker clusters, audio-visual association with a set of video streams is performed so that the video or approximate spatio-temporal location of a speaker can be found from multiple cameras. We carried out experiments showing whether it is possible to associate *all* participants to their audio source correctly in a batch manner and how the performance degrades as the length of the meeting is shortened. As the window size gets smaller, the likelihood of more than 1 person speaking within the same time interval is greatly reduced so we finally carried out experiments on selecting and evaluating whether *just the speaker* was associated with the correct video stream.
(c-d) Extraction of visual activity features : The video features themselves are computed in the compressed domain to take advantage of processing that is already required for the video compression process. Using these features, it is possible to do some spatial video-processing in order to identify the locations of two participants in video streams. We try using different sets of cameras to both represent and localize speakers in the meeting. Finally, to improve localization performance, we tried creating a binary representation of each person's visual activity, which generated a cleaner signal than the original raw features used.

1.3 The Augmented MultiParty Interaction (AMI) Corpus

Fig. 1.2 All available views in the data set.

One of the largest corpora of meeting room data has been recorded by the Augmented MultiParty Interaction (AMI) corpus which was created out of a European Union funded project [11]. This initiative generated a corpus that contains both 100 hours of audio-visual data and annotations from semantically low-level features, such as who is speaking to, more semantically meaningful concepts, such as dialogue acts or who is looking at whom. In each meeting, four participants were grouped together and were asked to design a remote control device over a series of

sessions. Each person was assigned a role such as “Project Manager”, “Marketing Expert”, or “Industrial Designer”. A microphone array and four cameras were set in the center of the room. Side and rear cameras were also mounted to capture different angles of the meeting room and its participants, as shown in Figure 1.2.

Each camera captures the visual activity of a single seated participant, who is assigned a seat at the start of each meeting session. Participants are requested not to change seats during the session. No other people enter or leave the meeting during the session so there are always only 4 interacting participants. Each person also wore a headset and a lapel microphone. A plan view of the meeting room is shown in Figure 1.3.

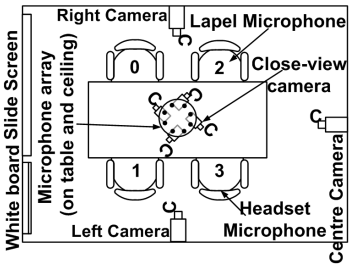


Fig. 1.3 Plan of the experimental meeting room.

1.4 Audio Speaker Diarization

1.4.1 Traditional Offline Speaker Diarization

As previously explained in Section 1.1, the goal of speaker diarization is answering the question “who spoke when?”. The following section outlines the traditional audio-only speaker diarization approach as shown in Figure 1.4.

Feature Extraction Wiener filtering is first performed on the audio channel for noise reduction. The HTK toolkit[1] is used to convert the audio stream into 19-dimensional Mel-Frequency Cepstral Coefficients (MFCCs) which are used as features for diarization. A frame period of 10 ms with an analysis window of 30 ms is used in the feature extraction.

Speech/Non-Speech Detection The speech/non-speech segmentation [64] proceeds in three steps. At each step, feature vectors consisting of 12 MFCC components, their deltas and delta-deltas (approximations of first and second order derivatives), and zero-crossings are used.

In the first step, an initial segmentation is created by running the Viterbi algorithm on a Hidden Markov Model (HMM) with Gaussian Mixture Model (GMM) emissions that have been trained on Dutch broadcast news data to segment speech and silence. In the second step, the non-speech regions are split into two clusters:

[1] http://htk.eng.cam.ac.uk/

regions with low energy and regions with high energy. A new and separate GMM is then trained on each of the two new clusters and on the speech region. The number of Gaussians used in the GMM is increased iteratively and re-segmentation is performed in each iteration. The model that is trained on audio with high energy levels is added to the non-speech model to capture non-speech-like sounds such as music, slamming doors, paper rustling, etc. In the final step, the speech model is compared to all other models using the Bayesian Information Criterion (BIC). If the BIC score is positive, the models are added to the speech model.

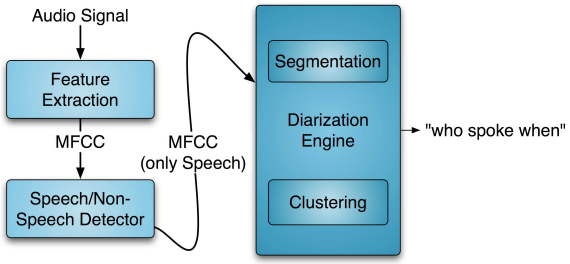


Fig. 1.4 Block diagram illustrating the traditional speaker diarization approach: As described in Section 1.4, an agglomerative clustering approach combines speaker segmentation and clustering in one step.

Speaker Segmentation and Clustering In the segmentation and clustering stage of speaker diarization, an initial segmentation is first generated by randomly partitioning the audio track into k segments of the same length. k is chosen to be much larger than the assumed number of speakers in the audio track. For meetings data, we use $k = 16$. The procedure for segmenting the audio data takes the following steps:

1. Train a set of GMMs for each initial cluster.
2. Re-segmentation: Run a Viterbi decoder using the current set of GMMs to segment the audio track.
3. Re-training: Retrain the models using the current segmentation as input.
4. Select the closest pair of clusters and merge them. This is done by going over all possible pairs of clusters, and computing the difference between the sum of the Bayesian Information Criterion (BIC) scores of each of the models and the BIC score of a new GMM trained on the merged cluster pair. The clusters from the pair with the largest positive difference are merged, the new GMM is used and the algorithm repeats from the re-segmentation step.
5. If no pair with a positive difference is found, the algorithm stops, otherwise the algorithm repeats from step 2.

A more detailed description can be found in [**?**].

The result of the algorithm consist of a segmentation of the audio track with n clusters and an audio GMM for each cluster, where n is assumed to be the number of speakers.

The computational load of such a system can be decomposed into three components: (1) find the best merge pair and merge; (2) model re-training and re-alignment; (3) other costs. After profiling the run-time distribution of an existing speaker diarization system, we find that the BIC score calculation takes 62 % of the total run-time.

Analyzing how the best merge hypothesis is found, the reason for the high cost of the BIC score calculation can be identified. Let D_a and D_b represent the data belonging to cluster a and cluster b, which are modeled by θ_a and θ_b, respectively. D represents the data after merging a and b, i.e. $D = D_a \cup D_b$, which is parameterized by θ. The Merge Score (MS) is calculated as Eq. (1.1) [1]:

$$MS(\theta_a, \theta_b) = \log p(D|\theta) - (\log p(D_a|\theta_a) + \log p(D_b|\theta_b)) \tag{1.1}$$

For each merge hypothesis a and b, a new GMM (θ) needs to be trained. When the system is configured to use more initial clusters, which is preferable for better initial cluster purity, the computational load can become prohibitive.

The speaker diarization output consists of meta-data describing speech segments in terms of starting time, ending time, and speaker cluster name. This output is usually evaluated against manually annotated ground truth segments. A dynamic programming procedure is used to find the optimal one-to-one mapping between the hypothesis and the ground truth segments so that the total overlap between the reference speaker and the corresponding mapped hypothesized speaker cluster is maximized. The difference is expressed as Diarization Error Rate (DER), which is defined by NIST[2]. The DER can be decomposed into three components: misses (speaker in reference, but not in hypothesis), false alarms (speaker in hypothesis, but not in reference), and speaker errors (mapped reference is not the same as hypothesized speaker).

This Speaker Diarization System has competed in the NIST evaluations of the past several years and established itself well among state-of-the-art systems[3].

The current official score is 21.74 % DER for the single-microphone case (RT07 evaluation set). This error is composed of 6.8 % speech/non-speech error and 14.9 % speaker clustering error. The total speaker error includes all incorrectly classified segments, including overlapped speech. NIST distinguishes between recordings with multiple distant microphones (MDM) and recordings with one single distant microphone (SDM). In the case of MDM, beam-forming is typically performed to produce a single channel out of all available ones.

For our approach, the various experimental conditions that we used can be categorized into a single distant microphone case and a individual close-talk microphone. 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.

[2] http://nist.gov/speech/tests/rt/rt2004/fall

[3] NIST rules prohibit publication of results other than our own. Please refer to the NIST website for further information: http://www.nist.gov/speech/tests/rt/rt2007

1.4.2 Online Speaker Diarization

Our first goal is to segment live-recorded audio into speaker-homogeneous regions to answer the question 'who is speaking now?'. For the system to work live and on-line, the question must be answered on intervals of captured audio that are as small as possible, and performed in at least real-time. The on-line speaker diarization system has been described in detail in [62] and has two steps: (*i*) training and (*ii*) recognition, which will be described in more detail in the subsequent sections.

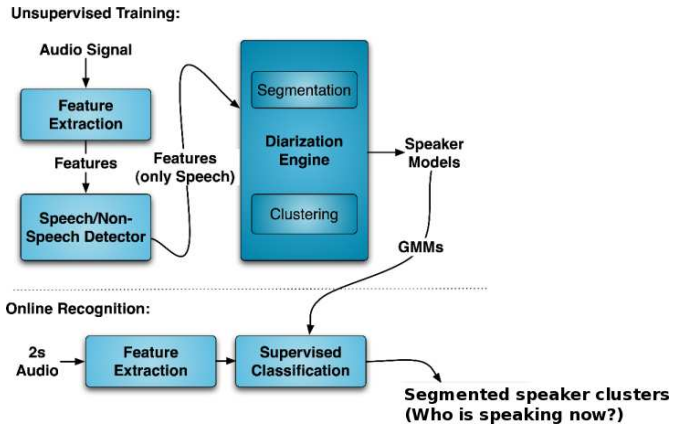


Fig. 1.5 Summary of the on-line audio diarization algorithm.

Unsupervised Bootstrapping of Speaker Models To bootstrap the creation of models, we use the speaker diarization system proposed by Wooters et al. [64] which was presented in Section 1.4.1 in the first meeting of each session. This also results in an estimation of the number of speakers and their associated speaker models. Once models have been created, they are added to the pool of speaker models and can be reused for all subsequent meetings. The speaker diarization system used for training is explained as follows. **Speaker Recognition** In recognition mode, the system records and processes chunks of audio as follows. First, Cepstral Mean Subtraction (CMS) is implemented to reduce stationary channel effects [56]. While some speaker-dependent information is lost, according to our experiments performed, the major part of the discriminant information remains in the temporally varying signal. In the classification step, the likelihood for each audio frame is computed against each set of Gaussian Mixtures obtained in the training step. From our previous experiments on larger meeting corpora, [62], we decided to use two-second chunks of audio. This introduces a latency of about 2.2 seconds after the person has started talking (recording 200 audio frames at 10 ms intervals plus a processing time of 0.1 $\times$ real time).

The decision on whether a segment belongs to a certain speaker or the non-speech model is reached using majority vote on the likelihoods of an audio frame belonging to a GMM. If the audio segment is classified as speech, we compare the winning speaker model against the second best model by computing the likelihood ratio. We use this as an indicator of the confidence level. In our experiments, we assume that

there are speaker models for all possible speakers so we used the highest confidence level to indicate the most likely speaker. For a more realistic case, it is possible to apply a threshold to the confidence level to detect an unknown speaker but this currently requires manual intervention.

A Note on Model Order Selection Offline audio speaker diarization can lead to more clusters than speakers since the method is data-driven and therefore cluster merging stops depending on whether the BIC score is improved or worsened by merging two candidate clusters. Due to the robustness of our on-line speaker diarization algorithm, while more clusters than participants can be generated in the offline training phase, in the on-line stage, noisy or extraneous clusters have much lower likelihoods, so they are never selected as likely speaker models. We found in our experiments that the number of recognized clusters and that of actual participants were always equal.

It is also important to note that the data that we use includes overlapping speech. These periods are automatically ignored when the speaker models are generated to ensure they remain as clean as possible. Work has been carried out to address overlapping speech in offline diarization systems but involve a second pass over the diarized audio signal, which would not be feasible for an on-line and real-time system [8].

1.4.3 Summary of the Diarization Performance

As described earlier, the output of a speaker diarization system consists of metadata describing speech segments in terms of start and end times, and speaker cluster labels. NIST provides a measurement tool that uses a dynamic programming procedure to find the optimal one-to-one mapping between the hypothesis and the ground truth segments so that the total overlap between the reference speaker and the corresponding mapped hypothesized speaker cluster is maximized. The difference is expressed as Diarization Error Rate, which is also defined by NIST[4]. The Diarization Error Rate (DER) can be decomposed into three components: misses (speaker in reference, but not in hypothesis), false alarms (speaker in hypothesis, but not in reference), and speaker errors (mapped reference is not the same as hypothesized speaker). It is expressed as a percentage relative to the total length of the meeting.

To characterize the algorithm under increasingly noisy input conditions, 3 different sources were used. Two signals were obtained by mixing the four individual headset microphones (MH) or lapel microphones (ML) using a direct summation. Also a real far-field case (F) where a single microphone from the array on the table was used. Table 1.1 shows the results for the on-line audio diarization system where the average, best and worse performances are shown for 12 meeting sessions that were used. As expected, one can observe a decrease in performance as the SNR decreases. It was interesting to observe a high variation in performance where in one case the error rate fell to 4.53% for the mixed headset condition. If we observe the variation in performance more closely, as shown in Figure 1.6, we see that there is one particular meeting session which has a consistently better performance than the

[4] http://nist.gov/speech/tests/rt/rt2004/fall

Input	Offline Results			Online Results		
Video Methods	F (21dB)	ML (22dB)	MH (31dB)	F (21dB)	ML (22dB)	MH (31dB)
Average DER(%)	33.16	36.35	36.16	18.26	26.18	28.57

Table 1.1 Diarization results in terms of the Diarization Error Rate (DER) using both offline and on-line methods. Note that the offline results were computed using meetings of 5-minute length while the on-line results were bootstrapped using longer meetings but speaker models were produced from just 60s of speech from each person. Results are also presented using different microphone sources where the associated signal to noise ratio for each source is shown in brackets.

rest. This is because in this meeting, everyone stays seated (and therefore maintains equidistance from the far-field microphone). In addition, the meeting is mostly a discussion and there is little use of the other equipment in the room such as the slide screen or white board. In contrast, meeting IS1006d is one of the worst performing meetings because people are often presenting at the whiteboard or slide screen. It is also interesting to observe that while the relative performance when using the far-field and headset microphones remain fairly consistent (the far-field case always performs worse), the mixed lapel condition does not. This could be explained by additional noise generated by shifting of the body or touching the microphone by accident, particularly when participants were moving around the meeting room.

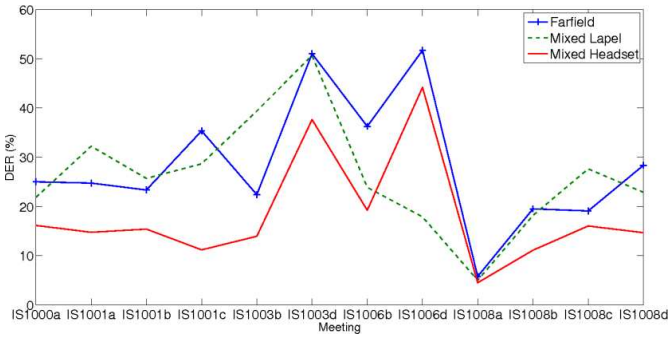


Fig. 1.6 Comparison of the on-line speaker diarization performance across different input conditions and over the different meetings that were considered.

1.5 Extracting Computationally Efficient Video Features

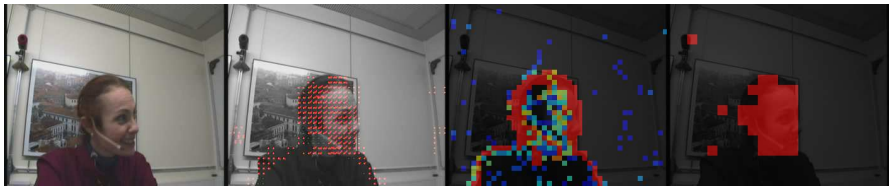
With the increased need for recording and storing video data, many modern day video cameras have hardware to encode the signal at the source. In order to capture visual activity efficiently, we leverage the fact that meeting videos are already in compressed form so that we can extract visual activity features at a much lower computational cost.

These features are generated from compressed-domain information such as motion vectors and block discrete-cosine transform coefficients that are accessible with almost zero cost from compressed video [63]. As compared to extracting similar

higher resolution pixel-based features such as optical flow, compressed domain features are much faster to extract, with a run-time reduction of 95% [67].

Video streams that have been compressed using MPEG4 encoding contains a collection of group-of-picture (GOP) which is structured with an Intra-coded frame or I-frame while the rest are predicted frames or P-frames. Figure 1.7 summarizes the various compressed domain features which can be extracted cheaply from compressed video as the *motion vector magnitude* (see Figure 1.7(b)) and the *residual coding bitrate* (see Figure 1.7(c)) to estimate visual activity level. Motion vectors, illustrated in Figure 1.7(d), are generated from motion compensation during video encoding; for each source block that is encoded in a predictive fashion, its motion vectors indicate which predictor block from the reference frame (in this case the previous frame for our compressed video data) is to be used. Typically, a predictor block is highly correlated with the source block and hence similar to the block to be encoded. Therefore, motion vectors are usually a good approximation of optical flow, which in turn is a proxy for the underlying motion of objects in the video [14].

After motion compensation, the DCT-transform coefficients of the residual signal (the difference between the block to be encoded and its prediction from the reference frame) are quantized and entropy coded. The *residual coding bitrate*, illustrated in Figure 1.7(c), is the number of bits used to encode this transformed residual signal. While the motion vector captures gross block translation, it fails to fully account for non-rigid motion such as lips moving. On the other hand, the residual coding bitrate is able to capture the level of such motion, since a temporal change that is not well-modeled by the block translational model will result in a residual with higher energy, and hence require more bits to entropy encode.



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

Fig. 1.7 Compressed domain video feature extraction. (a) Original image, (b) Motion vectors, (c) Residual coding bit-rate, (d) skin-colored regions.

1.5.1 Estimating personal activity levels in the compressed domain

Even when personal close-view cameras are used, the distance from the camera causes scale and pose issues, as shown in some example shots in Figure 1.8. By averaging activity measures over detected skin-color blocks, we hope to mitigate some of these issues. Therefore we implement a block-level skin-color detector that works mostly in the compressed domain which can detect head and hand regions as illustrated in Figure 1.7. This is also useful for detecting when each meeting participant is in view. To do this, we use a GMM to model the distribution of chrominance coefficients [41] in the YUV color-space. Specifically, we model the chrominance

Fig. 1.8 Possible pose variations and ambiguities captured from the video streams.

coefficients, (U,V), as a mixture of Gaussians, where each Gaussian component is assumed to have a diagonal covariance matrix. In the Intra-frames of the video, we compute the likelihood of observed chrominance DCT DC coefficients according to the GMM and threshold it to determine skin-color blocks. Skin blocks in the Inter-frames are inferred by using motion vector information to propagate skin-color blocks through the duration of the group-of-picture (GOP).

We threshold the number of skin-colored blocks in the close-up view to detect when a participant is seated. If a participant is not detected in an image frame of the close-up video stream, he is assumed to be presenting at the projection screen, which is a reasonable assumption in the meeting data. Since they are assumed to be presenting at the slide screen or whiteboard, they are more likely to be active and also speaking. Therefore, a simple assumption was to set periods where the person was detected as not seated, to the maximum value seen so far. While this is a simple rule, it was found to be effective in previous experiments [31].

1.5.2 Finding personal head and hand activity levels

While previous work has concentrated on extracting personal visual activity from gross head motion, here we go a step further by trying to understand how head and hand motion might play a part in human discourse, at a holistic level. The importance of this can be highlighted in Figure 1.9 where we observe three seconds of a meeting discussion. There are four participants in the discussion, in the configuration shown in Figure 1.3. Here we see just two participants where the person on the right is speaking. The top two rows of Figure 1.9 shows a breakdown of the gross head and hand motion that is observed for the two observed meeting participants, illustrated in the bottom row of the figure. To illustrate the change in motion over time more clearly, the average motion vector magnitudes over the head and hand skin regions are shown (further details about how these are calculated will be provided in the remainder of this section). The visual head and hand activity for the silent participant on the left is shown in grey while the speaker's visual activity is shown in black. The bottom two rows of the figure shows some key image frames within the three second interval where the person on the right is speaking. She starts off addressing those on the other side of the table and then directly addresses the participant to the left half way through the observed interval. When he realizes that he is being addressed directly, he moves his head to face her directly, but then lowers it again when attention is shifted away from him. In terms of hand motion, we see that the speaker is the only person of the two that moves during this interval. Note that in this paper, we describe head motion to be observed from skin-color

regions, which captures visual activity inside the face as well as some translations and deformations of the face region.

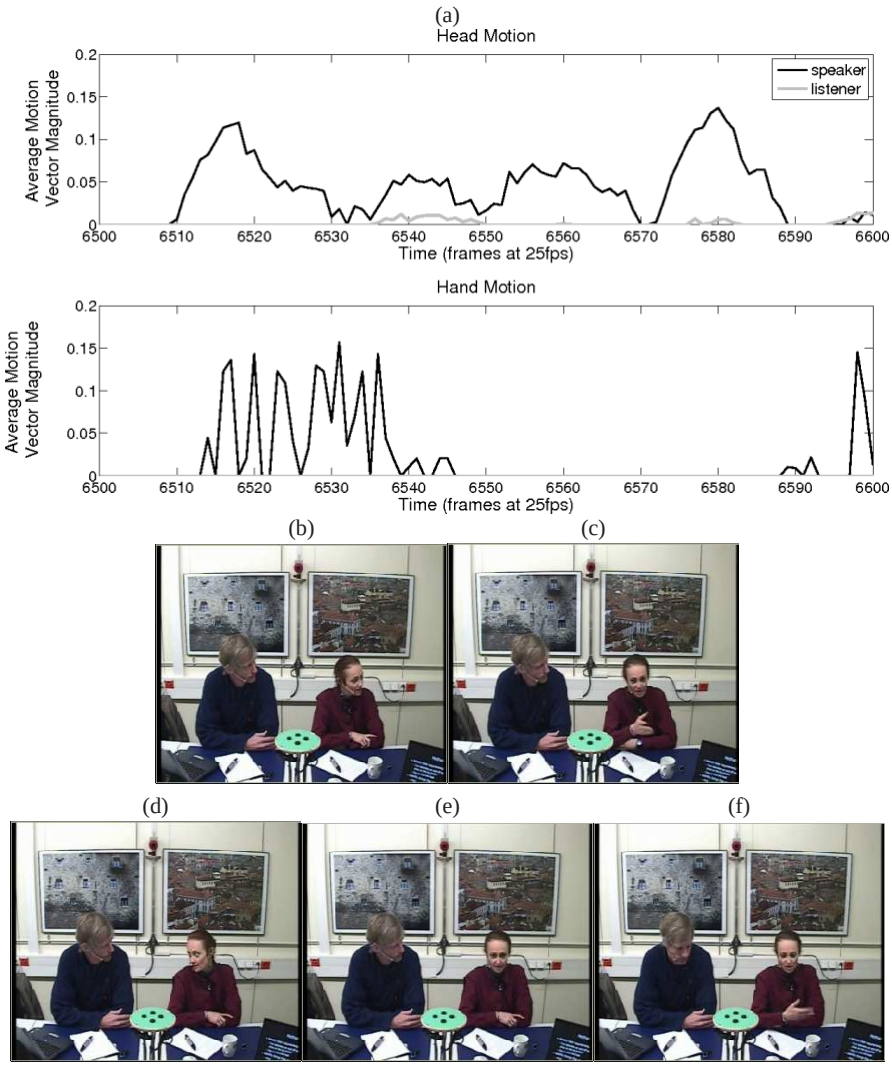


Fig. 1.9 Illustration of the difference in head and hand motion between speaker and listener. The black lines show the head and hand motion of the speaker while those in grey show the motion of the listener. The two rows below shows key image frames from this 3s interval where the person on the right is speaking the entire time.

The example in Figure 1.9 shows that a speaker and an attentive listener can have very different behavior types if we simply observe the head and hand motion separately. It is also interesting to observe that partial occlusion of one of the hands does

not affect the discrimination between the roles of these two meeting participants. Of course, the data is not always as clean and depends on how involved the participants were. Note also that the motion vector magnitudes were shown for illustrative purposes only; in our experiments, we use the residual coding bit-rate, which we found to produce better results since it tends to smooth out large and fast variations in the visual activity, and can also detect small motions from the lips if they are visible.

The features extraction method described in Section 1.5.1 were for gross body motion, and can include both head and hand motion where the hands are only sporadically visible in the close-up views (see bottom row of Figure 1.2). Therefore, we focus on extracting the desired features from the side views (see image L and R of the top row of Figure 1.2) where two people's head and hands are captured.

We first need to find the boundary between the two persons in each side view. The method we employ was inspired by the work of Jaimes on studying body postures of office workers [33]. For each image frame, we construct a horizontal profile of the number of detected skin-color blocks in each column, as shown by the accumulated profile at the bottom of the image in Figure 1.10. Suppose $S(x,y)$ is an indicator function of skin color blocks for the (x,y) block in the image frame. The horizontal profile is simply $S_h(x) = \sum_y S(x,y)$. Since we expect the horizontal location of each person's head to result in a strong peak in $S_h(x)$, we use a K-means clustering algorithm (with $K = 2$) to find the locations of the two peaks. To ensure continuity between image frames, K-means is initialized with the locations of the peaks from the previous image frame. The boundary is simply the mid-point between the two peaks. Once the left and right region of each camera-view is separated, we treated the two portions of the image frame as two video streams, representing the individual visual activity of each person in the same way as described in Section 1.5.1.

Next, we needed to find the boundary between the head and hands for each person. This time, for each person (i.e. the left half or right half of the view, separated by the estimated boundary), we constructed a vertical profile of the number of detected skin-color blocks in each row as shown in Figure 1.10. Again, since we expect the vertical location of the head and hands to result in strong peaks in the vertical profile, we use a K-means algorithm to find the two peaks. As before, K-means is initialized with the locations of the peaks from the previous image frame, and the boundary between the head and hands is just the mid-point. Note that the vertical profile is only considered below a certain height to remove spurious detections of skin color in the background.

Now, we can compute head and hands activity levels using the same approach as in Section 1.5.1, except that the area of interest is the estimated quadrant of the side-view that contains the subject of interest, i.e. left person's head, left person's hands, right person's head and right person's hands.

We evaluated the boundary estimation described above on one meeting session, where bounding boxes of speakers' heads had been annotated. The error rate of finding the boundary between two persons was 0.4%, where an error is defined as the estimated boundary not cleanly separating the bounding boxes of the two persons. The error rate of finding the boundary between the head and hands is 0.5%, where an error is defined as the estimated boundary not being below the head bounding

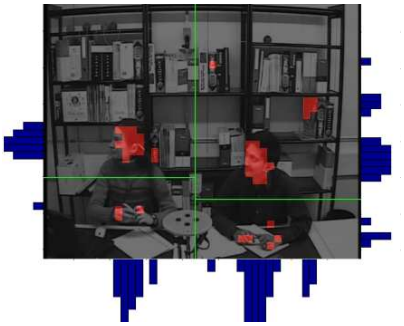


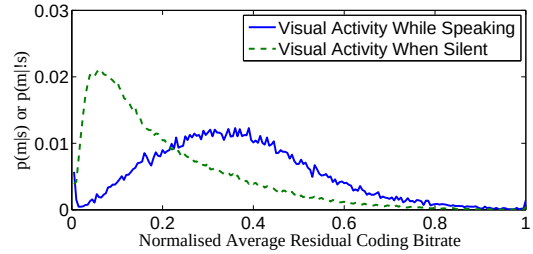
Fig. 1.10 Example of the horizontal and vertical profiles of the skin blocks and the located boundaries between the two people and their respective head and hand regions. The accumulated horizontal of the skin-color blocks is shown at the bottom of the example snap-shot. The vertical profiles of the skin-color blocks for each corresponding person is shown to the left and right of the image frame. The detected skin color regions are highlighted in red and the estimated boundaries using the horizontal and vertical profiles, are shown in green.

boxes of the respective person. We found that errors occurred mostly when the hands touched the face or moved above the shoulders or when a person reached across the table to their neighbor's table area. From this 2-camera set-up, four different personal activity features were generated; head activity; hand activity; the average activity of the head and hand blobs; and the maximum of the average head and average hand motion after the features were normalized.

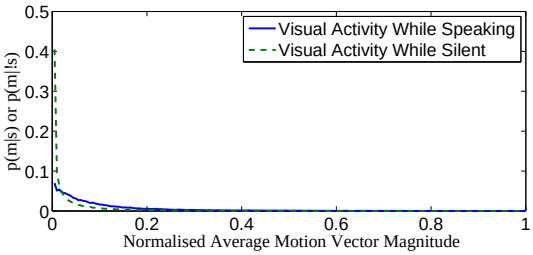
1.5.3 Estimating speakers using video only

From previous experiments, we have found that speech and the visual activity of the speaker are better correlated over long-term intervals [32, 31]. We know that people who move are not necessarily talking but we know that people who talk will tend to move. This is further illustrated by the distributions in Figure 1.11(a) where we see accumulated histograms of the distribution of visual activity as measured using the residual coding bit-rate with the close-up cameras, when people were seated and speaking or silent. This shows that people who talk tend to move more but that people who are silent can sometimes move a lot too. As mentioned in Section 1.5.1, when a person is detected as standing, their visual activity level is set to the highest value for that person that has been observed so far. Note also that previously [32] we found that using the motion vectors to associate audio and video streams led to worse performance. This is further illustrated in Figure 1.11(b) where the same distributions as (a) are shown but using the average motion vector magnitude instead.

To estimate the speaker based on observing the meeting participant with the most motion, it is important to first normalize the visual activity features for each person. The normalization allows us to compare the speaking and silent behavior of each participant in the meeting across all participants. For our meetings, there are no participants who remain inactive for the entire meeting therefore, we apply the normalization assuming that all participants will be relatively engaged in the meeting



(a)



(b)

Fig. 1.11 The accumulated visual activity histograms over our data-set during speaking (dashed line) and silence (solid line) for all participants for both the average residual coding bit-rate features in (a) and also the average motion vector magnitude in (b).

activities. Since the method is on-line, the normalization needed to be adaptive and so each new visual activity value was divided by the maximum value that was observed until that point.

Once the values have been normalized, each person's visual activity stream is considered to be comparable across individuals. Using this assumption and also that we know that speakers tend to move more than listeners, binary versions of each person's speaking activity was estimated. This was done by making the person who had the highest visual activity over the previous time window the estimated speaker, as described in Algorithm 1. This makes the same assumption as the speaker diarization system, that the speech is not overlapped, though in reality overlapping regions of speech exist in our test data, and are usually the periods in which correct estimates are more difficult to make. As discussed previously, it would have been interesting to account for cases of overlapping speech but previous work has shown that this would require a second pass over the data in order to find regions where the likelihood of a particular person speaking becomes much lower than during periods of clean speech [8].

1.6 Associating Speaker Clusters with Video Channels

To begin with, let us consider how well speech and audio streams can be associated together if clean audio signals are used. We used speaker segmentations from the audio signal taken from personal headset microphones as a simple automated speaker segmentation method. These were associated with the two real-valued visual activity features using the residual coding bit-rate or motion vector magnitudes. The

foreach *p in Participants* **do**
  $Votes[p] = 0$;
end
foreach *t in Window* **do**
  $i = argmax_p(VisualActivity[t, p]),\ \forall p \in Participants$;
  $Votes[i] = Votes[i] + 1$;
end
$j = argmax_p(Votes[p]),\ \forall p \in Participants$;
BinaryVisualActivity[j]=1;

Algorithm 1: Estimating speakers using visual activity only.

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.

For each pair-wise combination of speaking and visual activity channels, their corresponding normalized correlation was calculated. We then matched the channels by using an ordered one-to-one mapping based on associating the best correlated channels first. Figure 1.12 shows the algorithm in more detail.

(a) Quantifying the distance between audio-visual streams: the pair-wise correlation between each video, v_i, and audio stream, a_j, is calculated:-

$$\rho_{v_i,a_j} = \frac{\sum_{t=0}^{T} v(t) \cdot a(t)}{\sum_{t=0}^{T} v(t) \sum_{t=0}^{T} a(t)},\ \forall\{i, j\} \tag{1.2}$$

where T is the total length of the meeting and in our experiments t indexes the feature value at each frame. For our experiments, the frame rate used was 5 frames per second.

(b) Selecting the closest audio-visual streams : the pair of audio and video streams with the highest correlation are selected.

(c) Selection of the next closest audio-visual streams : the next best correlated pair of audio and video streams is selected.

(d) Full assignment of audio and video streams: step (c) is repeated until all audio-visual streams are associated.

Since the association is performed on a meeting basis, it is important to evaluate the performance similarly. Three evaluation criteria are used, to observe the difficulty in associating more channels correctly in each meeting. Hard (EvH), medium (EvM), and soft (EvS), criteria are used which assigns respectively a score of 1 for each meeting only when all, at least two, or at least one of the pairs of associated audio and visual streams is correct for each meeting. We refrain from evaluating on a participant basis since the meeting-based ordered mapping procedure, by definition, discriminates pairs that are easier to distinguish, as a means of improving the association from noisier channels which may have less observable activity.

The proportion of correctly associated meetings using both visual activity feature types are shown in Table 1.2 below. Correlating the headset segmentations and Residue visual activity channels performed best. Also, it was also encouraging to see that even for the hard evaluation strategy, the performance remained high for this case.

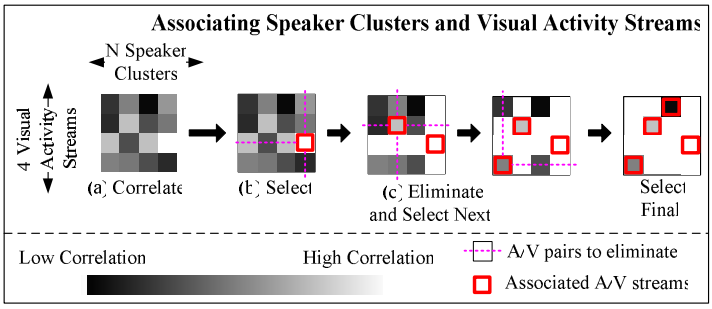


Fig. 1.12 Greedy Algorithm for ordered and discriminative pairwise associations between audio and video streams. (i) All pairwise combinations of the audio and video streams are correlated. (ii) The pair with the highest correlation is associated first and then eliminated from the correlation matrix.

	EvS	*EvM*	*EvH*
Residue	1.00	1.00	0.90
Vector	1.00	0.95	0.81

Table 1.2 Proportion of correctly associated meetings using speech segmentations generated from individual headset microphones that were then associated with visual activity from the close-view cameras. *EvH*: Hard evaluation strategy where all audio-visual streams in the meeting must be associated correctly; *EvM* Medium evaluation strategy where at least 2 of the audio-visual streams in the meeting must be associated correctly; *EvS* Soft evaluation strategy where at least 1 of the audio-visual streams in the meeting must be associated correctly.

For the on-line association method, the association method described above was modified so that after all streams were associated within a 2s sliding window. Then, only the person who spoke for the longest time was assigned their associated video stream for that window.

1.7 Audio-visual Clustering Results

Speaker localization experiments were run on the same meeting data that was used in the previous section. The outputs from the on-line speaker diarization were used as a reference to determine which video stream contained the relevant speaker. As described in Section 1.5 the visual activity of each individual could be represented by a number of features. These are summarized in Table 1.3. In addition, a binary feature can be derived from each of these using the method described in Section 1.5.3.

4 close-up cameras	Head Close-up			
2 mid-view cameras	Head+Hands	Head	Hands	Max(Head,Hands)

Table 1.3 Summary of video features that were used.

For the 4-camera and 2-camera case, the location of each stream was known so evaluation was straightforward. For the 2-camera case, it was assumed that each

half of the frame would be treated as a single stream, leading to 4 possible video candidates. An analysis window of 2s was used with a 40ms shift.

1.7.1 Using Raw Visual Activity

As an initial extension to the method presented in the previous subsection, we applied the on-line association algorithm to the real-valued average residual coding bit-rate in the 5 video forms described in Table 1.3. The results are summarized in Table 1.4 where evaluation was done using the same scoring as for the on-line diarization. Rather than comparing the speaker reference with the speaker clusters, which was done for the speaker diarization evaluation, we compare the speaker reference with the estimated video stream labels. For clarity, we refer to this measure as the association error rate (AER) but the mechanics of the performance measure are the same as the DER. We see that the error is quite high in all cases but note that the results are still better than random, where the error would be closer to 80% since the associated video could be one of the 4 participants or none of them. Comparing the performance more carefully across the different input audio conditions, we see that there is again a slight improvement in performance when the mixed headset signal is used rather than the far-field microphone. Comparing across the different video features that were tried, using the mean residual coding bit-rate for the estimated hand regions from the 2-camera set-up for each person gave the best results, but there was not a significant difference between the best and worse average results.

Input Video Methods	**Input audio conditions**					
	F (21dB)		**ML (22dB)**		**MH (31dB)**	
	AER (%)	**(Min)**	**AER (%)**	**(Min)**	**AER (%)**	**(Min)**
Head(Closeup)	68.39	(64.92)	68.42	(65.45)	68.04	(64.82)
Max(Head,Hands)	68.05	(62.75)	67.91	(62.09)	68	(60.62)
Heads	68.1	(64.25)	67.84	(63.79)	67.98	(63.03)
Head+Hands	67.67	(61.54)	67.58	(61.87)	67.54	(61.31)
Hands	67.92	(61.41)	67.65	(61.29)	67.64	(61.13)

Table 1.4 Audio-visual speaker localization with the real-valued average residual coding bit-rate for each person, using the different video feature methods. The signal-to-noise ratio for each audio type is shown in brackets for each input audio source. The results show the average AER over all the meetings for each experimental condition where the bracketed number shows the lowest AER that was achieved.

1.7.2 Using estimates of speaking activity from video

We then conducted similar experiments with each video feature type replaced by it's binarized version using the method described in Section 1.5.3. These binarized video streams were then associated with the relevant audio stream as described in Section 1.6. The results are summarized in terms of AER again in Table 1.5. Here we see a significant increase in performance when these binarized visual activity values are used. This indicates that our hypothesis that people who talk tend to move more is quite successful at finding speakers from video only. Overall, the best

speaker and video association performance was observed when the motion from the close-up cameras was used. This is not surprising since the head is represented at a higher resolution and therefore lip motion is better captured. It is encouraging, to see that even when using the 2-camera set-up, where the size of the heads was about half of those in the close-view cameras, the performance is slightly worse but still comparable. Of the 2-camera features, the one using head activity alone gave the best average performance but the best performance for any session used the Max(Head,Hands) feature. This indicates that hand motion can still be effective for discriminating speakers from listeners and is complementary to head motion. The worse average AER of the Max(Head,Hands) case compared to the Heads is likely to be due to how much body motion was attributed to meeting activities such as using a laptop, writing or manipulating the remote control prototype they were designing.

Input Video Methods	**Input audio conditions**					
	F (21dB)		**ML (22dB)**		**MH (31dB)**	
	AER (%)	**(Min)**	**AER (%)**	**(Min)**	**AER (%)**	**(Min)**
Head(Close-up)	41.89	(20.19)	41.91	(19.71)	41.55	(19.71)
Max(Head,Hands)	42.38	(22.24)	42.82	(22.37)	42.83	(22.39)
Heads	42.3	(26.27)	42.75	(26.42)	42.62	(26.4)
Head+Hands	46	(33.3)	46.83	(33.41)	46.24	(33.31)
Hands	53.83	(34.48)	54.79	(34.55)	54.18	(34.67)

Table 1.5 Audio-visual speaker localization results using binary estimates of speaking status from each person's visual activity. The signal-to-noise ratio for each audio type is shown in brackets for each input audio source. The results show the average AER for each experimental condition and the accompanying bracketed number shows the minimum AER that was achieved from one of the 12 sessions that were used.

Since the AER is not a widely used performance measure, in multi-modal processing tasks, we also provide the average precision, recall and F-measure when using the far-field microphone and binary estimates of speaking activity in Table 1.6. Here the boldened values show the best achieved performance for a single meeting while the number on the left shows the average. Using these measures, similar differences in performance are observed, although here, using the maximum of the head and hand motion appears to give the best overall performance for the 2-camera case. Again, the 4-camera case performs the best. It is also interesting to observe that the head-only and the Max(Head,Hands) features perform similarly while the Head+Hands and hands-only features perform similarly badly compared to the rest. This indicates that for both listeners and speakers, observing head motion is more discriminative in most situations. However, the success of the feature which takes the maximum of the head and hand motion indicates that the head and hand features should be treated independently since they are complementary.

From the results we have presented, it seems that using the binary estimates of speaking activity from video is effective. However, the performance is not as high as estimating speakers from the audio alone. We can observe the locations of failure modes by looking more closely at an example meeting, which is shown in Fig-

Input Video Methods	Prec.	Recall	F-meas.
Head(Close-up)	52.74 **72.93**	41.64 **62.53**	44.72 **66.18**
Max(Heads,Hands)	50.64 **68.62**	41.58 **62.26**	43.59 **63.1**
Head	51.01 **66.41**	41.95 **58.18**	43.93 **60.2**
Head+Hands	39.63 **56.51**	34.17 **54.21**	34.68 **49.44**
Hands	37.17 **56.91**	31.33 **48.12**	31.64 **43.28**

Table 1.6 Summary of the average precision, recall and F-measure for the different video feature types. Results for using the far-field microphone are shown and the binary estimates of speaking status from visual activity. For each video feature, the highest performance is shown boldened.

ure 1.13. Here the binary segmentations of the estimated speaker are shown using the association method described in Section 1.6 (first row); the binary estimates of speaking activity from video (second row); and the speaker clusters generated from the on-line speaker diarization algorithm (third row). The final row shows the ground truth speaker segmentations. We can see that there are occasions (e.g. between 150s-200s and 600s-700s) when the binary estimates of speaking activity fail since the person who moves the most isn't talking. This is not surprising since there is still a considerable overlap observed in the speaker and listener activity shown in Figure 1.11 previously. Furthermore, we observed that there are occasions where non-speakers were involved in other activities while someone was speaking (e.g. working on a laptop). However, there are also observed cases where speaker diarization fails and the speaker estimates from video was successful (between 350s-450s). The failure in the speaker diarization could be caused by speaker models being confused due to either short utterances or because the speaker models were only generated from 60s of speech for each speaker in the training phase. This example of complementary failure modes suggests that combining the audio and video features at an earlier stage may also improve the speaker diarization performance.

1.8 Discussion

In this chapter, we have discussed off-line systems which can be used for post-processing of previously recorded data. However, audio-visual mining of the data could also happen in real-time. A system that can work on-line and in real-time is useful for remote meeting scenarios where subtle information about an interaction can be lost through transmission. These could relate to transmission failure of one or more modalities but could also be due to the inherent time delay between sending and receiving data. In terms of more complex immersion problems within the remote meeting scenario, it is also difficult for remote participants to know when to interrupt in a conversation or judge the mood or atmosphere of the group they are interacting with. For co-located meetings, live-recording and summary may be useful for a quick recap if someone missed what was said (e.g. a phone call interruption) but doesn't want to interrupt the conversation flow in order to catch up on information they missed. Aside from this, live processing also aids post-meeting browsing since a live capability could be used to enable live tagging of automatically segmented events such as how an issue on the agenda was received by other meeting participants. Of course, some of the tags could be substituted by automated

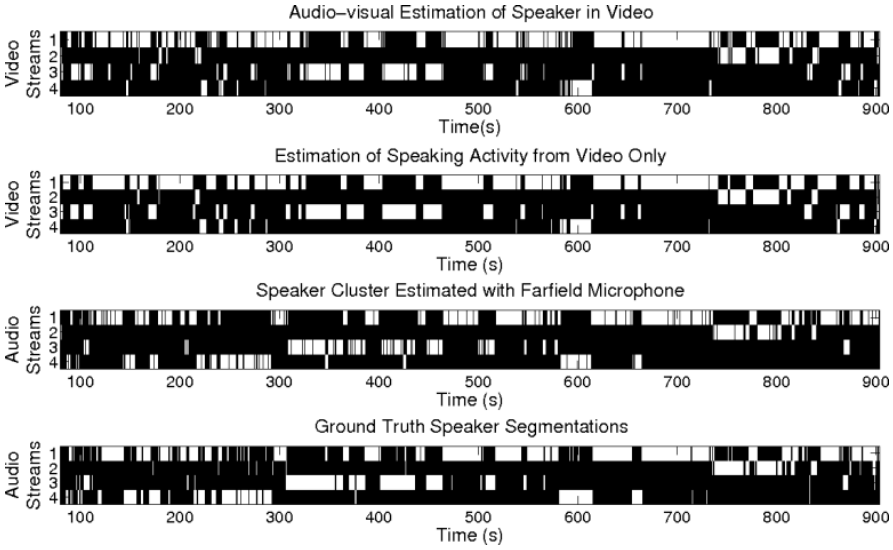


Fig. 1.13 Graphical comparison of different feature representations and estimates. White areas indicate either that someone is speaking. The first row shows the estimated associated video stream, given the diarized speaker clusters in the third row; The second row shows the estimate of speaker status from just the motion activity taken from the maximum of the head and hand motion; and the final row shows the ground-truth speaker segmentations.

event identification methods but when certain technologies are not available, tagging is an extremely useful way of labeling information. In particular, tagging has been used extensively for mining image data with the emergence of social networking sites where photos are organized amongst self-organized groups. It has been demonstrated that imagery data itself need not be used for mining the data if tags are available [44].

Moving away from the on-line and real-time problems, there are other ways in which the performance of the speaker diarization and audio-visual association task can be improved. In particular, while the approach presented in this chapter demonstrated a late fusion approach, given that we know that speech and body motion it correlated, there is also motivation to make the task into a speaker diarization and localization task by fusing the modalities early on in the clustering process. This is particularly interesting since clustering video data alone into speakers tends to require apriori knowledge of the number or participants. Of course, techniques such as face detection can be employed to identify the speakers but this may not be practical if the resolution of faces is in the video and non-frontal faces tend to be difficult to detect robustly. Research on fusing audio and visual features for speaker diarization or speaker localization as discussed in Section 1.1 has shown an improvement in performance over single-modality methods. However most work performs experiments on data where two talking heads are visible and remain relatively stationary with fully frontal faces. Few consider more global body movements [31, 32, 10, 59, 60]. Vajaria et al. [59, 60] was one of the first to use gross body movement for speaker

diarization and localization but suffer from the need to cluster spatially separated noisy visual features. Recently some preliminary success by using just a single camera and microphone [23] to perform speaker diarization where the audio and visual features are fused early on in the agglomerative clustering process. Results for the speaker diarization task show improvement, despite the low resolution of each participant in the captured video. In both cases, the correlation of speech and motion from different body parts was not considered for the diarization task. Also, finding a suitable way to evaluate the locations of speakers in the video in a similar way to the diarization performance is yet to be found.

With the success of multi-modal speaker diarization methods, it is clear that the trend is moving towards using multiple sensors and multiple modalities to solve data-mining problems, certainly in the domain of meeting analysis. The importance of multi-modal data mining when capturing human behavior is further emphasized since psychologically, both modalities are used differently when we communication socially and communicate very different messages. It is sometimes these differences and in particular unusual events which trigger memories for us about a particular conversation. It could be said that these are the events which are the most difficult to find again once they have been archived. This brings us to the application of estimating dominance, which was demonstrated at the end of this chapter. It showed that even with computationally efficient methods for clustering the data where the estimates of the raw outputs was degraded, the performance of the semantically higher level dominance task was not necessarily affected. This addresses some interesting questions about how the problem of the semantic gap should be addressed in data mining. From a cognitive perspective, perhaps we would expect that the verbal content of each speaker would need to be analyzed. However, experiments have shown that using speaking time alone, is quite robust, even if the estimates of the speaker turns are not as accurate. Given these results, one might ask the question of whether other semantically high-level behavioral types or affiliations can be characterized using equally simple features such as the excitement levels in a meeting [65], roles [68], or personality [52].

Ultimately, one could argue that to address the semantic gap in mining meeting data, we must start from the questions we ask ourselves when trying to search through meeting data such as in terms of what happened, what were the conclusions and how people interacted with each other. From a functional perspective, knowing the meeting agenda and the final outcomes are useful but from a social perspective knowing about the subtle non-verbal behavior tells us more about relationships between colleagues or clients. For example, knowing how a person usually behaves can help us to detect unusual behavior, which could be indications of stress, if for example the person has been delegated too much work. These are ultimately useful tools to ensure that teams in organizations work effectively and that staff are not overworked or under-utilized. From an individual perspective, there are those that argue that success is well correlated with “emotional intelligence” which is defined as the ability to monitor both one’s own and the other’s feelings and emotions in order to guide one’s thinking and actions [51]. Automatically estimating the feelings and emotions of others are topics of interest currently [65, 7]. In particular, recent

work on distinguishing real from fake facial expressions of pain has shown that automated systems perform significantly better than human observers [7]. Such research shows the potential of using machines to help us understand how we interact and in particular how this could potentially be used to help individuals in becoming more aware of social interactions around them. Ultimately, such knowledge should lead to more efficient team-working where perhaps the easiest failure mode in teams occurs through a break-down in communication between members.

Acknowledgments

This research was partly funded by the US VACE program, the EU project AMIDA (pub. AMIDA-103), the Swiss NCCR IM2, and the Institute for Infocomm Research, Singapore.

References

1. J. Ajmera and C. Wooters. A robust speaker clustering algorithm. In *Proc. IEEE Automatic Speech Recognition Understanding Workshop*, 2003.
2. M. Al-Hames, A. Dielmann, D. Gatica-Perez, S. Reiter, S. Renals, G. Rigoll, and D. Zhang. Multimodal integration for meeting group action segmentation and recognition. In *Proc. Multimodal Interaction and Related Machine Learning Algorithms Workshop (MLMI–05)*, pages 52–63. Springer, 2006.
3. X. Anguera. *Robust Speaker Diarization for Meetings*. PhD thesis, Technical University of Catalonia, Barcelona, Spain, December 2006.
4. X. Anguera, B. Peskin, and M. Aguilo. Robust speaker segmentation for meetings: The icsi-sri spring 2005 diarization system. In *Proc. of NIST MLMI Meeting Recognition Workshop, Edinburgh*, 2005.
5. D.F. Armstrong, W.C. Stokoe, and S. Wilcox. *Gesture and the Nature of Language*, chapter What is gesture?, pages 40–41. Cambridge University Press, 1995.
6. R.F. Bales. *Interaction process analysis: A method for the study of small groups*. Cambridge, Addison-Wesley, 1950.
7. M. Bartlett, G. Littlewort, E. Vural, K. Lee, M. Cetin, A. Ercil, and J. Movellan. Data mining spontaneous facial behavior with automatic expression coding. *In Verbal and Nonverbal Features of Human-Human and Human-Machine Interaction*, pages 1–20, 2008.
8. K. Boakye, B. Trueba-Hornero, O. Vinyals, and G. Friedland. Overlapped speech detection for improved speaker diarization in multiparty meetings. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4353–4356, 2008.
9. J.P. Campbell. Speaker recognition: a tutorial. *Proceedings of the IEEE*, pages 1437–1462, 1997.
10. N. Campbell and N. Suzuki. Working with Very Sparse Data to Detect Speaker and Listener Participation in a Meetings Corpus. In *Workshop Programme*, volume 10, May 2006.
11. J.C. 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 *Joint Workshop on Machine Learning and Multimodal Interaction (MLMI)*, 2005.
12. S. Chen and P. Gopalakrishnan. Speaker, environment and channel change detection and clustering via the bayesian information criterion. In *Proceedings of DARPA speech recognition workshop*, 1998.

13. T. Chen and R.R. Rao. Cross-modal Prediction in Audio-visual Communication. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 4, pages 2056–2059, 1996.
14. M. T. Coimbra and M. Davies. Approximating optical flow within the MPEG-2 compressed domain. *IEEE Transactions on Circuits and Systems for Video Technology*, 15(1):103–107, 2005.
15. W.S. Condon and W.D. Ogston. Sound film analysis of normal and pathological behavior patterns. *The Journal of Nervous and Mental Disease*, 143(4):338, 1966.
16. W.S. Condon and W.D. Osgton. Speech and body motion synchrony of the speaker-hearer. *The Perception of Language*, pages 150–184, 1971.
17. J.M.J.R. DABBS and RB Ruback. Dimensions of group process: Amount and structure of vocal interaction. *Advances in experimental social psychology*, 20:123–169, 1987.
18. P. Delacourt and C.J. Wellekens. Distbic: A speaker-based segmentation for audio data indexing. *Speech Communication: Special Issue in Accessing Information in Spoken Audio*, pages 111–126, 2000.
19. J.F. Dovidio and S.L. Ellyson. Decoding visual dominance: Attributions of power based on relative percentages of looking while speaking and looking while listening. *Social Psychology Quarterly*, 45(2):106–113, 1982.
20. S. Favre, H. Salamin, J. Dines, and A. Vinciarelli. Role recognition in multiparty recordings using social affiliation networks and discrete distributions. In *IMCI '08: Proceedings of the 10th international conference on Multimodal interfaces*, pages 29–36, New York, NY, USA, 2008. ACM.
21. J. W. Fisher and T. Darrell. Speaker association with signal-level audiovisual fusion. *IEEE Transactions on Multimedia*, 6(3):406–413, 2004.
22. J. W. Fisher, T. Darrell, W. T. Freeman, and P. A. Viola. Learning joint statistical models for audio-visual fusion and segregation. In *Conference on Neural Information Processing Systems (NIPS)*, pages 772–778, 2000.
23. G. Friedland, H. Hung, and C. Yeo. Multi-modal speaker diarization of real-world meetings using compressed-domain video features. In *Proceedings of IEEE ICASSP*, April 2009.
24. G. Friedland, O. Vinyals, Y. Huang, and Christian Mueller. Prosodic and other Long-Term Features for Speaker Diarization. *Transactions on Audio, Speech and Language Processing*, 17, May 2009.
25. J.B. Gatewood and R. Rosenwein. Interactional synchrony: Genuine or spurious? A critique of recent research. *Journal of Nonverbal Behavior*, 6(1):12–29, 1981.
26. H. Gish and M. Schmidt. Text-independent speaker identification. *IEEE Signal Processing Magazine*, pages 18–32, 1994.
27. U. Hadar, TJ Steiner, and F. Clifford Rose. Head movement during listening turns in conversation. *Journal of Nonverbal Behavior*, 9(4):214–228, 1985.
28. U. Hadar, TJ Steiner, EC Grant, and F. Clifford Rose. The timing of shifts of head postures during conversation. *Human Movement Science*, 3:237–245, 1984.
29. T. Hain, L. Burget, J. Dines, G. Garau, M. Karafiat, D. van Leeuwen, M. Lincoln, and V. Wan. The 2007 AMI(DA) system for meeting transcription. In *Multimodal Technologies for Perception of Humans*, volume 4625 of *Lecture Notes in Computer Science*, pages 414–428. Springer-Verlag, 2008. International Evaluation Workshops CLEAR 2007 and RT 2007, Baltimore, MD, USA, May 8-11, 2007, Revised Selected Papers.
30. J. A. Harrigan. Listener's body movements and speaking turns. *Communication Research*, 12:233–250, 1985.
31. H.Hung and G. Friedland. Towards audio-visual on-line diarization of participants in group meetings. In *Workshop on Multi-camera and Multi-modal Sensor Fusion Algorithms and Applications in conjunction with ECCV*, Marseille, France, October 2008.
32. H. Hung, Y. Huang, C. Yeo, and Daniel Gatica-Perez. Associating audio-visual activity cues in a dominance estimation framework. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshop on Human Communicative Behavior*, Alaska, 2008.

33. A. Jaimes. Posture and activity silhouettes for self-reporting, interruption management, and attentive interfaces. In *International Conference on Intelligent User Interfaces*, pages 24–31, New York, NY, USA, 2006. ACM.
34. D. B. Jayagopi, S. O. Ba, J.-M. Odobez, and D. Gatica-Perez. Predicting two facets of social verticality in meetings from five-minute time slices and nonverbal cues. In *Proceedings - ICMI 2008*, 2008.
35. D. B. Jayagopi, H. Hung, C. Yeo, and D. Gatica-Perez. Modeling dominance in group conversations using nonverbal activity cues. *IEEE Transactions on Audio, Speech and Language Processing*, 2008.
36. B. Juang and L. Rabiner. A probabilistic distance measure for hidden markov models, 1985.
37. A. Kendon. *Conducting Interaction: Patterns of Behavior in Focused Encounters*. Cambridge University Press, 1990.
38. A. Kendon. Some considerations for a theory of language origins. *Man, New Series*, 26(2):199–221, 1991.
39. H.-G. Kim and T. Sikora. Hybrid speaker-based segmentation system using model-level clustering. In *Proceedings of the IEEE International Conference on Audio and Speech Signal Processing*, 2005.
40. L. 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*, 27(3):305–317, 2005.
41. S. J. McKenna, S. Gong, and Y. Raja. Modelling facial colour and identity with Gaussian mixtures. *Pattern Recognition*, 31(12):1883–1892, 1998.
42. D. McNeill. *Hand and Mind. What Gestures Reveal About Thought. Chicago: Univ*. Chicago Press, 1992.
43. D. McNeill. *Language and Gesture*. Cambridge University Press New York, 2000.
44. R.-A. Negoescu and D. Gatica-Perez. Topickr: Flickr groups and users reloaded. In *MM '08: Proc. of the 16th ACM Intl. Conf. on Multimedia*, Vancouver, Canada, October 2008. ACM.
45. H. Ning, M. Liu, H. Tang, and T. Huang. A spectral clustering approach to speaker diarization, the ninth international conference on spoken language processing. In *Proceedings of Interspeech*, 2006.
46. H. J. Nock, G. Iyengar, and Chalapathy Neti. Speaker localisation using audio-visual synchrony: An empirical study. In *ACM International Conference on Image and Video Retrieval*, pages 488–499, 2003.
47. A. Noulas and B. J. A. Krose. On-line multi-modal speaker diarization. In *International Conference on Multimodal Interfaces*, pages 350–357, New York, USA, 2007. ACM.
48. 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.
49. K. C. H. Parker. Speaking turns in small group interaction: A context-sensitive event sequence model. *Journal of Personality and Social Psychology*, 54(6):965–971, 1988.
50. E. K. Patterson, S. Gurbuz, Z. Tufekci, and J. N. Gowdy. CUAVE: A new audio-visual database for multimodal human-computer interface research. In *International Conference on Acoustics, Speech, and Signal Processing*, pages 2017–2020, 2002.
51. K. V. Petrides, N. Frederickson, and A. Furnham. The role of trait emotional intelligence in academic performance and deviant behavior at school. *Personality and Individual Differences*, 36(2):277–293, January 2004.
52. F. Pianesi, N. Mana, A. Cappelletti, B. Lepri, and M. Zancanaro. Multimodal recognition of personality traits in social interactions. In *International Conference on Multimodal interfaces*, pages 53–60, New York, NY, USA, 2008. ACM.
53. R. Rao and T. Chen. Exploiting audio-visual correlation in coding of talking head sequences. *International Picture Coding Symposium*, March 1996.
54. S. Reiter, B. Schuller, and G. Rigoll. Hidden conditional random fields for meeting segmentation. In *2007 IEEE International Conference on Multimedia and Expo*, pages 639–642, 2007.

55. D. A. Reynolds and P. Torres-Carrasquillo. Approaches and applications of audio diarization. In *Proc. of International Conference on Audio and Speech Signal Processing*, 2005.
56. Douglas A. Reynolds. Speaker identification and verification using Gaussian mixture speaker models. *Speech Communication*, 17(1-2):91–108, 1995.
57. M.R. Siracusa and J.W. Fisher. Dynamic dependency tests for audio-visual speaker association. In *International Conference on Acoustics, Speech and Signal Processing*, April 2007.
58. S. Tamura, K. Iwano, and S. FURUI. Multi-Modal Speech Recognition Using Optical-Flow Analysis for Lip Images. *Real World Speech Processing*, 2004.
59. H. Vajaria, T. Islam, S. Sarkar, R. Sankar, and R. Kasturi. Audio segmentation and speaker localization in meeting videos. *International Conference on Pattern Recognition*, 2:1150–1153, 2006.
60. H Vajaria, S. Sarkar, and R. Kasturi. Exploring co-occurence between speech and body movement for audio-guided video localization. *IEEE Transactions on Circuits and Systems for Video Technology*, 18:1608–1617, Nov 2008.
61. O. Vinyals and G. Friedland. Modulation Spectrogram Features for Speaker Diarization. *Proceedings of Interspeech*, pages 630–633, September 2008.
62. O. Vinyals and G. Friedland. Towards semantic analysis of conversations: A system for the live identification of speakers in meetings. In *Proceedings of IEEE International Conference on Semantic Computing*, pages 456–459, August 2008.
63. H. Wang, A. Divakaran, A. Vetro, S.F. Chang, and H. Sun. Survey of compressed-domain features used in audio-visual indexing and analysis. *Journal of Visual Communication and Image Representation*, 14(2):150–183, 2003.
64. C. Wooters and M. Huijbregts. The ICSI RT07s speaker diarization system. In *Proceedings of the NIST RT07 Meeting Recognition Evaluation Workshop*. Springer, 2007.
65. B. Wrede and E. Shriberg. Spotting "hot spots" in meetings: Human judgments and prosodic cues. In *in Proc. Eurospeech*, pages 2805–2808, 2003.
66. S. J. Wrigley, G. J. Brown, V. Wan, and S. Renals. Speech and crosstalk detection in multichannel audio. *IEEE Trans. on Speech and Audio Processing*, 13:84–91, 2005.
67. C. Yeo and K. Ramchandran. Compressed domain video processing of meetings for activity estimation in dominance classification and slide transition detection. Technical Report UCB/EECS-2008-79, EECS Department, University of California, Berkeley, Jun 2008.
68. M. Zancanaro, B. Lepri, and F. Pianesi. Automatic detection of group functional roles in face to face interactions. In *International Conference on Multimodal interfaces*, pages 28–34, New York, NY, USA, 2006. ACM.
69. C. Zhang, P. Yin, Y. Rui, R. Cutler, and P. Viola. Boosting-Based Multimodal Speaker Detection for Distributed Meetings. *IEEE International Workshop on Multimedia Signal Processing (MMSP) 2006*, 2006.