# Multimodal Integration for Meeting Group Action Segmentation and Recognition

Marc Al-Hames [1] Alfred Dielmannéthoz [2]
Daniel Gatica-Perez [3] Stephan Reiter [1]
Steve Renalséthoz [2] Dong Zhang [3] [a]

IDIAP–RR 05-31

June. 2005

SUBMITTED FOR PUBLICATION

Dalle Molle Institute for Perceptual Artificial Intelligence • P.O.Box 592 • Martigny • Valais • Switzerland

phone +41 − 27 − 721 77 11
fax +41 − 27 − 721 77 12
e-mail secretariat@idiap.ch
internet http://www.idiap.ch

[1] Institute for Human-Machine-Communication, Technische Universität München Arcisstr. 21, 80290 Munich, Germany
[2] Centre for Speech Technology Research, University of Edinburgh 2 Buccleuch Place, Edinburgh EH8 9LW, UK
[3] IDIAP Research Institute, CP 592, 1920 Martigny, Switzerland

[a] Authors listed in alphabetical order.

IDIAP Research Report 05-31

# Multimodal Integration for Meeting Group Action Segmentation and Recognition

Marc Al-Hames Alfred Dielmannéthoz Daniel Gatica-Perez
Stephan Reiter Steve Renalséthoz Dong Zhang [1]

June. 2005

SUBMITTED FOR PUBLICATION

# Abstract

We address the problem of segmentation and recognition of sequences of multimodal human interactions in meetings. These interactions can be seen as a rough structure of a meeting, and can be used either as input for a meeting browser or as a first step towards a higher semantic analysis of the meeting. A common lexicon of multimodal group meeting actions, a shared meeting data set, and a common evaluation procedure enable us to compare the different approaches. We compare three different multimodal feature sets and four modelling infrastructures: a higher semantic feature approach, multi-layer HMMs, a multi-stream DBN, as well as a multi-stream mixed-state DBN for disturbed data.

## 1 Introduction

Recordings of multi-party meetings are useful to recall important pieces of information (decisions, key-points, milestones, etc.), and eventually share it with people who were not able to attend those meetings. Unfortunately, watching raw audio-video recordings is tedious. Therefore an automatic approach to extract high-level information could facilitate this task.

In this paper we address the problem of recognising sequences of human interaction patterns in meetings, with the goal of structuring them in semantic terms. Our aim is to discover repetitive patterns into natural group interactions and associate them with a lexicon of meeting actions or phases (such as discussions or monologues). The detected sequence of meeting actions will provide a relevant summary of the meeting structure. The investigated patterns are inherently group-based (involving multiple simultaneous participants), and multimodal (as captured by cameras and microphones).

Automatic modelling of human interactions from low-level multimodal signals is an interesting topic for both theoretical and practical reasons. First, from the theoretical point of view, modelling multichannel multimodal sequences provides a particular challenging task for machine learning techniques. Secondly, from the application point of view, automatic meeting analysis could add value to the raw data for browsing and retrieval purposes.

Starting from a common lexicon of meeting actions (section 2) and sharing the same meeting data-set (section 3), each site (TUM, IDIAP and UEDIN) has selected a specific feature set (section 4) and proposed relevant models (section 5). Then a common evaluation metric (section 6) has been adopted in order to compare several experimental setups (section 7).

## 2 Action Lexicon

Two sets of meeting actions have been defined. The first set (lexicon 1, defined in [8]) includes eight meeting actions, like discussion, monologue, or presentation. The monologue action is further distinguished by the person actually holding the monologue (e.g. monologue 1 is meeting participant one speaking). The second set (lexicon 2, defined in [15]) comprehends the full first set, but also has combinations of two parallel actions (like a presentation and note-taking). The second set includes fourteen group actions. Both sets and a brief description are shown in table 1.

## 3 Meeting Data Set

We used a public corpus of 59 five-minute, four-participant scripted meetings [8]. The recordings took place at IDIAP in an instrumented meeting room equipped with cameras and microphones[1]. Video has been recorded using 3 fixed cameras. Two cameras capture a frontal view of the meeting participants, and the third camera captures the white-board and the projector screen. Audio was recorded using lapel microphones attached to participants, and an eight-microphone array placed in the centre of the table.

[1] This corpus is publicly available from http://mmm.idiap.ch/

Table 1: Group action lexicon 1 and 2

| Action | Lexicon | Description |
|---|---|---|
| Discussion | lexicon 1 and 2 | most participants engaged in conversations |
| Monologue | lexicon 1 and 2 | one participant speaking continuously without interruption |
| Monologue+ Note-taking | contained only in lexicon 2 | one participant speaking continuously others taking notes |
| Note-taking | lexicon 1 and 2 | most participants taking notes |
| Presentation | lexicon 1 and 2 | one participant presenting using the projector screen |
| Presentation+ Note-taking | contained only in lexicon 2 | one participant presenting using projector screen, others taking notes |
| White-board | lexicon 1 and 2 | one participant speaking using the white-board |
| White-board+ Note-taking | contained only in lexicon 2 | one participant speaking using white-board, others taking notes |

# 4 Features

The investigated individual actions are multimodal, we therefore use different audio-visual features. They are distinguished between *person-specific* AV features and *group-level* AV features. The former are extracted from individual participants. The latter are extracted from the white-board and projector screen regions. Furthermore we use a small set of lexical features. The features are described in the next sections, for details please refer to the indicated literature.

From the large number of available features each site has chosen a set, used to train and evaluate their models. The complete list of features, and the three different sets IDIAP, TUM, UEDIN are listed in table 2.

## 4.1 Audio features

**MFCC:** For each of the speakers four MFC coefficients and the energy were extracted from the lapel-microphones. This results in a 20-dimensional vector $\vec{x}_S(t)$ containing speaker-dependent information.

**A binary speech and silence segmentation** (BSP) for each of the six locations in the meeting room was extracted with the SRP-PHAT measure [8] from the microphone array. This results in a six-dimensional discrete vector $\vec{x}_{BSP}(t)$ containing position dependent information.

**Prosodic features** are based on a denoised and stylised version of the intonation contour, an estimate of the syllabic rate of speech and the energy [5]. These acoustic features comprise a 12 dimensional feature vector (3 features for each of the 4 speakers).

**Speaker activity features** rely on the active speaker locations evaluated using a sound source localisation process based on a microphone array [8]. A 216 element feature vector resulted from all the $6^3$ possible products of the 6 most probable speaker locations (four seats and two presentation positions) during the most recent three frames [5]. A speaker activity feature vector at time $t$ thus gives a local sample of the speaker interaction pattern in the meeting at around time $t$.

**Further audio features:** From the microphone array signals, we first compute a speech activity measure (SRP-PHAT). Three acoustic features, namely energy, pitch and speaking rate, were estimated on speech segments, zeroing silence segments. We used the SIFT algorithm to extract pitch, and a combination of estimators to extract speaking rate [8].

Table 2: Audio, visual and semantic features, and the resulting three feature sets.

| | | **Description** | IDIAP | TUM | UEDIN |
|---|---|---|---|---|---|
| Person-Specific Features | Visual | head vertical centroid | X | | |
| | | head eccentricity | X | | |
| | | right hand horizontal centroid | X | | |
| | | right hand angle | X | | |
| | | right hand eccentricity | X | | |
| | | head and hand motion | X | | |
| | | global motion features from each seat | | X | |
| | Audio | SRP-PHAT from each seat | X | | |
| | | speech relative pitch | X | | X |
| | | speech energy | X | X | X |
| | | speech rate | X | | X |
| | | 4 MFCC coefficients | | X | |
| | | binary speech and silence segmentation | | X | |
| | Semantic | individual gestures | | X | |
| | | talking activity | | X | |
| Group Features | Visual | mean difference from white-board | X | | |
| | | mean difference from projector screen | X | | |
| | | global motion features from whiteboard | | X | |
| | | global motion features from projector screen | | X | |
| | Audio | SRP-PHAT from white-board | X | | |
| | | SRP-PHAT from projector screen | X | | |
| | | speaker activity features | | | X |
| | | binary speech from white-board | | X | |
| | | binary speech from projector screen | | X | |

## 4.2 Global motion visual features

In the meeting room the four persons are expected to be at one of six different locations: one of four chairs, the whiteboard, or at a presentation position. For each location $L$ in the meeting room a difference image sequence $I_d^L(x, y)$ is calculated by subtracting the pixel values of two subsequent frames from the video stream. Then seven global motion features [16] are derived from the image sequence: The centre of motion is calculated for the x- and y-direction, the changes in motion are used to express the dynamics of movements. Furthermore the mean absolute deviation of the pixels relative to the centre of motion is computed. Finally the intensity of motion is calculated from the average absolute value of the motion distribution. These seven features are concatenated for each time step in the location dependent motion vector. Concatenating the motion vectors from each of the six positions leads to the final visual feature vector that describes the overall motion in the meeting room with 42 features.

## 4.3 Skin-colour blob visual features

Visual features derived from head and hands skin-colour blobs were extracted from the three cameras. For the two cameras looking at people, visual features extracted consist of head vertical centroid position and eccentricity, hand horizontal centroid position, eccentricity, and angle. The motion magnitude for head and hand blobs were also extracted. The average intensity of difference images computed by background subtraction was extracted from the third camera. All features were extracted at 5 frames per second, and the complete set of features is listed in table 2. For details refer to [15].

### 4.4 Semantic features

Originating from the low level features also features on a higher level have been extracted. For each of the six locations in the meeting room the talking activity has been detected using results from [7]. Further individual gestures of each participant have been detected using the gesture recogniser from [16]. The possible features were all normalised to the length of the meeting event to provide the relative duration of this particular feature. From all available events only those that are highly discriminative were chosen which resulted in a nine dimensional feature vector.

## 5 Models for Group Action Segmentation and Recognition

### 5.1 Meeting segmentation using semantic features

This approach combines the detection of the boundaries and classification of the segments in one step. The strategy is similar to that one used in the BIC-Algorithm [14]. Two connected windows with variable length are shifted over the time scale. Thereby the inner border is shifted from the left to the right in steps of one second and in each window the feature vector is classified by a low-level classifier. If there is a different result in the two windows, the inner border is considered a boundary of a meeting event. If no boundary is detected in the actual window, the whole window is enlarged and the inner border is again shifted from left to the right. Details can be found in [13].

### 5.2 Multi-stream mixed-state DBN for disturbed data

In real meetings the data can be disturbed in various ways: events like slamming of a door may mask the audio channel or background babble may appear; the visual channel can be (partly) masked by persons standing or walking in front of a camera. We therefore developed a novel approach for meeting event recognition, based on mixed-state DBNs, that can handle noise and occlusions in all channels [1, 2]. Mixed-state DBNs are an HMM coupled with a LDS, they have been applied to recognising human gestures in [10]. Here, this approach has been extended to a novel multi-stream DBN for meeting event recognition.

Each of the three observed features: microphone array (BSP), lapel microphone (MFCC) and the visual global motion stream (GM) is modelled in a separate stream. The streams correspond to a multi-stream HMM, where each stream has a separate representation for the features. However, the visual stream is connected to a LDS, resulting in a mixed-state DBN. Here the LDS is a Kalman filter, using information from all streams as driving input, to smooth the visual stream. With this filtering, movements are predicted based on the previous time-slice and on the state of the multi-stream HMM at the current time. Thus occlusions can be compensated with the information from all channels. Given an observation $O$ and the model parameters $E_j$ for the mixed-state DBN, the joint probability of the model is the product of the stream probabilities: $P(O, E_j) = P_B \cdot P_M \cdot P_G$. The model parameters are learned for each of the eight event classes $j$ with a variational learning EM-algorithm during the training phase. During the classification an unknown observation $O$ is presented to all models $E_j$. Then $P(O|E_j)$ is calculated for each model and $O$ is assigned to the class with the highest likelihood: $\mathrm{argmax}_{E_j \in E} P(O|E_j)$. Applying the Viterbi-algorithm to the model, leads to a meeting event segmentation framework. The mixed-state DBN can therefore easily be combined with other models presented in this work.

### 5.3 Multi-layer Hidden Markov Model

In this section we summarise the multi-layer HMM applied to group action recognition. For a detailed discussion, please refer to [15].

In the multi-layer HMM framework, we distinguish group actions (which belong to the whole set of participants, such as *discussion and presentation*) from individual actions (belonging to specific persons, such as *writing and speaking*). To recognise group actions, individual actions act as the bridge between group actions and low-level features, thus decomposing the problem in stages, and simplifying the complexity of the task.

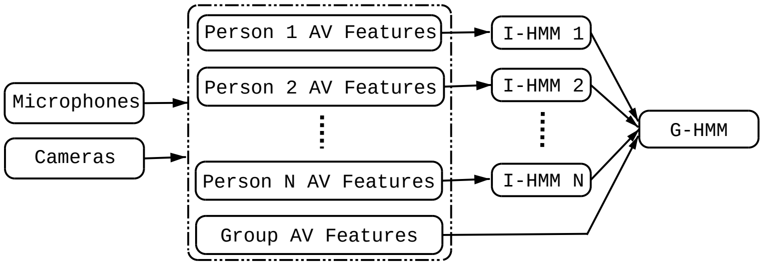

Figure 1: Multi-layer HMM on group action recognition.

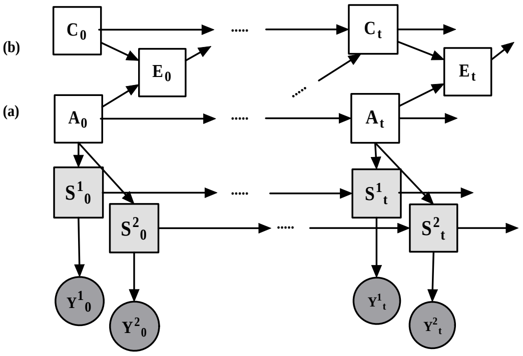

Figure 2: Multistream DBN model (a) enhanced with a “counter structure” (b); square nodes represent discrete hidden variables and circles must be intend as continuous observations

Let I-HMM denote the lower recognition layer (individual action), and G-HMM denote the upper layer (group action). I-HMM receives as input audio-visual (AV) features extracted from each participant, and outputs posterior probabilities of the individual actions given the current observations. In turn, G-HMM receives as input the output from I-HMM, and a set of group features, directly extracted from the raw streams, which are not associated to any particular individual. In the multi-layer HMM framework, each layer is trained independently, and can be substituted by any of the HMM variants that might capture better the characteristics of the data, more specifically asynchrony [3], or different noise conditions between the audio and visual streams [6]. The multi-layer HMM framework is summarised in figure 1.

Compared with a single-layer HMM, the layered approach has the following advantages, some of which were previously pointed out by [9]: (1) a single-layer HMM is defined on a possibly large observation space, which might face the problem of over-fitting with limited training data. It is important to notice that the amount of training data becomes an issue in meetings where data labelling is not a cheap task. In contrast, the layers in our approach are defined over small-dimensional observation spaces, resulting in more stable performance in cases of limited amount of training data. (2) The I-HMMs are person-independent, and in practice can be trained with much more data from different persons, as each meeting provides multiple individual streams of training data. Better generalisation performance can then be expected. (3) The G-HMMs are less sensitive to slight changes in the low-level features because their observations are the outputs of the individual action recognisers, which are expected to be well trained. (4) The two layers are trained independently. Thus, we can explore different HMM combination systems. In particular, we can replace the baseline I-HMMs with models that are more suitable for multi-modal asynchronous data sequences. The framework thus becomes simpler to understand, and amenable to improvements at each separate level.

### 5.4 Multistream DBN model

The DBN formalism allows the construction and development of a variety of models, starting from a simple HMM and extending to more sophisticated models (hierarchical HMMs, coupled HMMs, etc). With a small effort, DBNs are able to factorise the internal state space, organising it in a set of interconnected and specialised hidden variables.

Our multi-stream model (bottom of figure 2) exploits this principle in two ways: decomposing meeting actions into smaller logical units, and modelling parallel feature streams independently. We assume that a meeting action can be decomposed into a sequence of small units: meeting subactions. In accordance with this assumption the state space is decomposed into two levels of resolution: meeting actions (nodes $A$) and meeting subactions (nodes $S^F$). Note that the decomposition of meeting actions into meeting subactions is done automatically through the training process.

Feature sets derived from different modalities are usually governed by different laws, have different characteristic time-scales and highlight different aspects of the communicative process. Starting from this hypothesis we further subdivided the model state space according to the nature of features that are processed, modelling each feature stream independently (multistream approach). The resulting model has an independent substate node $S^F$ for each feature class $F$ , and integrates the information carried by each feature stream at a 'higher level' of the model structure (arcs between $A$ and $S^F$,$F = [1, n]$). Since the adopted *lexicon 1* (section 2) is composed by 8 meeting actions, the action node $A$ has a cardinality of 8. The cardinalities of the sub-action nodes $S$ are part of parameter set, and in our experiments we have chosen a value of 6 or 7.

The probability to remain in an HMM state corresponds to an inverse exponential [11]: a similar behaviour is displayed by the proposed model. This distribution is not well-matched to the behaviour of meeting action durations. Rather than adopting ad hoc solutions, such as action transition penalties, we preferred to improve the flexibility of state duration modelling, by enhancing the existing model with a counter structure (top of figure 2). The counter variable $C$, which is ideally incremented during each action transition, attempts to model the expected number of recognised actions. Action variables $A$ now also generate the hidden sequence of counter nodes $C$, together with the sequence of sub-action nodes $S$. Binary enabler variables $E$ have an interface role between action variables $A$ and counter nodes $C$.

This model presents several advantages over a simpler HMM in which features are "early integrated" into a single feature vector: feature classes are processed independently according to their nature; more freedom is allowed in the state space partitioning and in the optimisation of the sub-state space assigned to each feature class; knowledge from different streams is integrated together at an higher level of the model structure; etc. Unfortunately all these advantages, and the improved accuracy that can be achieved, are balanced by an increased model size, and therefore by an increased computational complexity.

## 6 Performance Measures

Since group meeting actions are high level symbols and their boundaries are extremely vague. In order to evaluate results of the segmentation and recognition task we used the Action Error Rate, a metric that privileges the recognition of the correct action sequence, rather than the precise temporal boundaries. AER is defined as the sum of *insertion* (Ins), *deletion* (Del), and *substitution* (Subs) errors, divided by the total number of actions in the ground-truth:

$$\text{AER} = \frac{\text{Subs} + \text{Del} + \text{Ins}}{\text{Total Actions}} \times 100\% \tag{1}$$

Measures based on *deletion* (Del) and *insertion* (Ins) and *substitution* (Subs) are also used to evaluate action recognition results.

Table 3: Segmentation results using the higher semantic feature approach (BN: Bayesian Network, GMM: Gaussian Mixture Models, MLP: Multilayer Perceptron Network, RBF: Radial Basis Network, SVM: Support Vector Machines). The columns denote the insertion rate, the deletion rate, the accuracy in seconds and the classification error rate (using lexicon 1 in Table 1).

| Classifier | Insertion (%) | Deletion (%) | Accuracy | Error (%) |
|---|---|---|---|---|
| BN | 14.7 | 6.22 | 7.93 | 39.0 |
| GMM | 24.7 | 2.33 | 10.8 | 41.4 |
| MLP | 8.61 | 1.67 | 6.33 | 32.4 |
| RBF | 6.89 | 3.00 | 5.66 | 31.6 |
| SVM | 17.7 | 0.83 | 9.08 | 35.7 |

# 7 Experiments and Discussions

## 7.1 Higher semantic feature approach

The results of the segmentation are shown in table 3 (BN: Bayesian Network, GMM: Gaussian Mixture Models, MLP: Multilayer Perceptron Network, RBF: Radial Basis Network, SVM: Support Vector Machines). Each row denotes the classifier that was used. The columns show the insertion rate (number of insertions in respect to all meeting events), the deletion rate (number of deletions in respect to all meeting events), the accuracy (mean absolute error) of the found segment boundaries in seconds and the recognition error rate. In all columns lower numbers denote better results. As can be seen from the tables, the results are quite variable and heavily depend on the used classifier. These results are comparable to the ones presented in [12]. With the integrated approach the best outcome is achieved by the radial basis network. Here the insertion rate is the lowest. The detected segment boundaries match pretty well with a deviation of only about five seconds to the original defined boundaries.

## 7.2 Multi-stream mixed-state DBN for disturbed data

To investigate the influence of disturbance to the recognition performance, the evaluation data was cluttered: the video data was occluded with a black bar covering one third of the image at different positions. The audio data from the lapel microphones and the microphone array was disturbed with a background-babble with 10dB SNR. 30 undisturbed videos were used for the training of the models. The remaining 30 unknown videos have been cluttered for the evaluation.

The novel DBN was compared to single-modal (audio and visual) HMMs, an early fusion HMM, and a multi-stream HMM. The DBN showed a significant improvement of the recognition rate for disturbed data. Compared to the baseline HMMs, the DBN reduced the recognition error by more than 1.5% (9% relative error reduction) for disturbed data. It may therefore be useful to combine this approach with the other models presented in this work, to improve the noise robustness. Please refer to [1, 2] for detailed recognition results, as well as a comprehensive description of the model.

## 7.3 Multi-layer hidden Markov model

Table 4 reports the performance in terms of action error rate (AER) for both multi-layer HMM and the single-layer HMM methods. Several configurations were compared, including audio-only, visual-only, early integration, multi-stream [6] and asynchronous HMMs [3]. We can see that (1) the multi-layer HMM approach always outperforms the single-layer one, (2) the use of AV features always outperforms the use of single modalities for both single-layer and multi-layer HMM, supporting the hypothesis that the group actions we defined are inherently multimodel, (3) the best I-HMM model is the asynchronous HMM, which suggests that some asynchrony exists for our task of group action recognition, and is actually well captured by the asynchronous HMM.

Table 4: AER (%) for single-layer and multi-layer HMM (using lexicon 2 in Table 1).

| **Method** | | AER (%) |
|---|---|---|
| Single-layer HMM | Visual only | 48.2 |
| | Audio only | 36.7 |
| | Early Integration | 23.7 |
| | Mutli-stream | 23.1 |
| | Asynchronous | 22.2 |
| Multi-layer HMM | Visual only | 42.4 |
| | Audio only | 32.3 |
| | Early Integration | 16.5 |
| | Multi-stream | 15.8 |
| | Asynchronous | 15.1 |

## 7.4 Multistream DBN model

All the experiments depicted in this section were conducted on 53 meetings (subset of the meeting corpus depicted in section 3) using the lexicon 1 of eight group actions. We implemented the proposed DBN models using the Graphical Models Toolkit (GMTK) [4], and the evaluation is performed using a leave-one-out cross-validation procedure.

Table 5 shows experimental results achieved using: an ergodic 11-states HMM, a multi-stream approach (section 5.4) with two feature streams, and the full counter enhanced multi-stream model. The base 2-stream approach has been tested in two different sub-action configurations: imposing $\left|S^1\right| = \left|S^2\right| = \{6\ or\ 7\}$. Therefore four experimental setups were investigated; and each setup has been tested with 3 different feature sets, leading to 12 independent experiments. The first feature configuration ("UEDIN") associates prosodic features and speaker activity features (section 4.1) respectively to the stream $S^1$ and to $S^2$. The feature configuration labelled as "IDIAP" makes use of the multimodal features extracted at IDIAP, representing audio related features (prosodic data and speaker localisation) through the observable node $Y^1$ and video related measures through $Y^2$. The last setup ("TUM") relies on two feature families extracted at the Technische Universität München: binary speech profiles derived from IDIAP speaker locations and video related global motion features; each of those has been assigned to an independent sub-action node. Note that in the HMM based experiment the only observable feature stream $Y$ has been obtained by merging together both the feature vectors $Y^1$ and $Y^2$. Considering only the results (of table 5) obtained within the UEDIN feature setup, it is clear that the simple HMM shows much higher error than any other multi-stream configuration. The adoption of a multistream based approach reduces the AER to less than 20%, providing the lowest AER (11%) when sub-action cardinalities are fixed to 7. UEDIN features seem to provide a higher accuracy if compared with IDIAP and TUM setups, but it is essential to remember that our DBN models have been optimised for the UEDIN features. In particular sub-action cardinalities have been intensively studied with our features, but it will be interesting to discover optimal values for IDIAP and TUM features too. Moreover overall performances achieved with the multistream approach are very similar (AER are always in the range from 26.7% to 11.0%), and all my be considered promising. The TUM setup seems to be the configuration for which switching from a HMM to a multistream DBN approach provides the greatest improvement in performance: the error rate decreases from 92.9% to 21.4%. If with the UEDIN feature set the adoption of a counter structure is not particularly effective, with IDIAP features the counter provides a significant AER reduction (from 26.7% to 24.9%). We are confident that further improvements with IDIAP features could be obtained by using more than 2 streams (such as the 3 multistream model adopted in [5]). Independently of the feature configuration, the best overall results are achieved with the multistream approach and a state space of 7 by 7 substates.

Table 5: AER (%) for an HMM, and for a multi-stream (2 streams) approach with and without the "counter structure"; the models have been tested with the 3 different feature sets (using lexicon 1)

| Model | Feature Set | Corr. | Sub. | Del. | Ins. | AER |
|---|---|---|---|---|---|---|
| HMM | UEDIN | 63.3 | 13.2 | 23.5 | 11.7 | 48.4 |
| | IDIAP | 62.6 | 19.9 | 17.4 | 24.2 | 61.6 |
| | TUM | 60.9 | 25.6 | 13.5 | 53.7 | 92.9 |
| 2 streams $\left(\left|S^F\right| = 6\right)$ | UEDIN | 86.1 | 5.7 | 8.2 | 3.2 | 17.1 |
| | IDIAP | 77.9 | 8.9 | 13.2 | 4.6 | 26.7 |
| | TUM | 85.4 | 9.3 | 5.3 | 6.8 | 21.4 |
| 2 streams $\left(\left|S^F\right| = 6\right)$ + counter | UEDIN | 85.8 | 7.5 | 6.8 | 4.6 | 18.9 |
| | IDIAP | 79.4 | 10.0 | 10.7 | 4.3 | 24.9 |
| | TUM | 85.1 | 5.7 | 9.3 | 6.4 | 21.4 |
| 2 streams $\left(\left|S^F\right| = 7\right)$ | UEDIN | 90.7 | 2.8 | 6.4 | 1.8 | 11.0 |
| | IDIAP | 86.5 | 7.8 | 5.7 | 3.2 | 16.7 |
| | TUM | 82.9 | 7.1 | 10.0 | 4.3 | 21.4 |

# 8 Conclusions

In this work, we have presented the joint efforts of the three institutes (TUM, IDIAP and UEDIN) towards structuring meetings into sequences of multimodal human interactions. A large number of different audio-visual features have been extracted from a common meeting data corpus. From this features, three multimodal sets have been chosen. Four different approaches towards automatic segmentation and classification of meetings into action units haven been proposed. We then deeply investigated the three feature sets, as well as the four different group action modelling frameworks:

The first approach from TUM exploits higher semantic features for structuring a meeting into group actions. It thereby uses an algorithm that is based on the idea of the Bayesian-Information-Criterion. The mixed-state DBN approach developed by TUM compensates for disturbances in both the visual and the audio channel. It is not a segmentation framework but can be integrated into the other approaches presented in this work to improve their robustness. The multi-layer Hidden Markov Model developed by IDIAP decomposes group actions as a two-layer process, one that models basic individual activities from low-level audio-visual features, and another one that models the group action (belonging to the whole set of participants). The multi-stream DBN model proposed by UEDIN operates an unsupervised subdivision of meeting actions into sequences of group sub-actions, processing multiple asynchronous feature streams independently, introducing also a model extension to improve state duration modelling.

All presented approaches have provided comparable good performances. The AER are already promising, but there is still space for further improvements both in the feature domain (i.e.: exploit more modalities) and in the model infrastructure. Therefore in the near future we are going to investigate combinations of the proposed systems to improve the AER and to exploit the complementary strengths of the different approaches.

### Acknowledgement

This work was partly supported by the European project M4 (MultiModal Meeting Manager) and European Union 6th FWP IST Integrated Project AMI (Augmented Multi-party Interaction, FP6-506811, publication AMI-87).

# References

[1] M. Al-Hames and G. Rigoll. A multi-modal graphical model for robust recognition of group actions in meetings from disturbed videos. In *Proc. IEEE ICIP*, Italy, 2005.

[2] M. Al-Hames and G. Rigoll. A multi-modal mixed-state dynamic Bayesian network for robust meeting event recognition from disturbed data. In *Proc. IEEE ICME*, 2005.

[3] S. Bengio. An asynchronous hidden markov model for audio-visual speech recognition. In S. Becker, S. Thrun, and K. Obermayer, editors, *Advances in NIPS 15*. MIT Press, 2003.

[4] J. Bilmes. Graphical models and automatic speech recognition. *Mathematical Foundations of Speech and Language Processing*, 2003.

[5] A. Dielmann and S. Renals. Multistream dynamic Bayesian network for meeting segmentation. *Lecture Notes in Computer Science*, 3361:76–86, 2005.

[6] S. Dupont and J. Luettin. Audio-visual speech modeling for continuous speech recognition. *IEEE Transactions on Multimedia*, 2(3):141–151, September 2000.

[7] G. Lathoud, I. A. McCowan, and J.-M. Odobez. Unsupervised Location-Based Segmentation of Multi-Party Speech. In *Proc. 2004 ICASSP-NIST Meeting Recognition Workshop*, 2004.

[8] I. McCowan, D. Gatica-Perez, S. Bengio, G. Lathoud, M. Barnard, and D. Zhang. Automatic analysis of multimodal group actions in meetings. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 27(3):305–317, 2005.

[9] N. Oliver, E. Horvitz, and A. Garg. Layered representations for learning and inferring office activity from multiple sensory channels. In *Proc. ICMI*, Pittsburgh, Oct. 2002.

[10] V. Pavlovic, B. Frey, and T.S. Huang. Time series classification using mixed-state dynamic Bayesian networks. In *Proc. IEEE CVPR*, 1999.

[11] L. R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proc. of the IEEE*, 2(77):257–286, 1989.

[12] S. Reiter and G. Rigoll. Segmentation and classification of meeting events using multiple classifier fusion and dynamic programming. In *Proc. IEEE ICPR*, pages 434–437, 2004.

[13] S. Reiter and G. Rigoll. Multimodal meeting analysis by segmentation and classification of meeting events based on a higher level semantic approach. In *Proc. IEEE ICASSP*, 2005.

[14] A. Tritschler and R.A. Gopinath. Improved speaker segmentation and segments clustering using the bayesian information criterion. In *Proc. EUROSPEECH '99*, 1999.

[15] D. Zhang, D. Gatica-Perez, S. Bengio, I. McCowan, and G. Lathoud. Modeling individual and group actions in meetings: a two-layer hmm framework. In *IEEE Workshop on Event Mining at the Conference on Computer Vision and Pattern Recognition (CVPR)*, 2004.

[16] M. Zobl, F. Wallhoff, and G. Rigoll. Action recognition in meeting scenarios using global motion features. In J. Ferryman, editor, *Proc. PETS-ICVS*, pages 32–36, 2003.