SEGMENTATION AND RECOGNITION OF MEETING EVENTS USING A TWO-LAYERED HMM AND A COMBINED MLP-HMM APPROACH

Stephan Reiter, Björn Schuller, and Gerhard Rigoll

Institute for Human-Machine-Communication
Technische Universität München
Arcisstr. 21, 80290 Munich, Germany
email: {reiter, schuller, rigoll}@ei.tum.de

ABSTRACT

Automatic segmentation and classification of recorded meetings provides a basis that enables effective browsing and querying in a meeting archive. Yet, robustness of today's approaches is often not reliable enough. We therefore strive to improve on this task by introduction of a hybrid approach combining the discriminative abilities of artificial neural nets and warping capabilities of hidden markov models. Dividing the task into two layers and defining a proper set of individual actions helps to cope with the problem of lack of data and overcomes conventional single-layered approaches. Extensive test runs on the public M4 Scripted Meeting Corpus prove the great performance gain applying our suggested novel approach compared to other similar methods.

1. INTRODUCTION

Automatic analysis of meetings has the potential to greatly reduce time and costs compared to human annotation. However, adequate robustness is yet to meet. Numerous research activities are therefore concerned with the development of reliable meeting recorder and browser systems: In the meeting project at ICSI [1], e.g., the main goal is to produce a transcript of the speech. At CMU the intention is to develop a meeting browser, which includes challenging tasks like speech transcription and summarization [2] and the multi-modal tracking of people throughout the meeting [3, 4]. In the European research project M4 the main concern is the construction of a demonstration system to enable structuring, browsing and querying of an archive of automatically analyzed meetings.

Due to the complex information flow of visual, acoustic and other information sources in meetings (e.g. from documents or projectors) the segmentation of a meeting in appropriate sections represents a very challenging pattern recognition task, which is of growing interest throughout an increasing number of research teams.

Goal of the described work here is, to automatically divide a meeting into segments with a length of several seconds, so called meeting events as *discussion*, *monologue* or *presentation*. A common approach is to present features in a sequential order as done in [5, 6, 7]. Thereby various standard techniques for pattern recognition are used. These include dynamic systems like Hidden Markov Models (HMM) as well as static approaches as Bayesian Networks, Multilayer Perceptrons (MLP) and Support Vector Machines (SVM) Also a layered approach using HMM at two different levels of granularity has been investigated recently [8]. However we propose a new layered approach not only using HMM but also using artificial neural networks.

The paper is organized as follows: Section 2 describes the database. In Section 3 our used features are described. Section 4 then gives an overview over the system structure and finally in Section 5 the results are presented.

2. MEETING CORPUS

Within our research we use the publicly available M4 Scripted Meeting Corpus, described in [5]. It consists of fully scripted meetings recorded in a Smart Meeting Room at IDIAP, equipped with fully synchronized multichannel audio and video recording facilities. Each of the recorded participants had a close-talk lapel microphone attached to his clothes. An additional microphone array was mounted on top of one center meeting table. Video signals were recorded onto separate digital video tape recorders by three television video cameras, providing PAL quality.

Each captured meeting consists of a set of predefined group actions in a fixed order defined in an according agenda. The appearing group actions are:

- Discussion (all participants engage in a discussion)
- Monologue (one participant speaks continuously without interruption)
- Note-taking (all participants write notes)
- Presentation (one participant at front of the room presents using the only projector screen)
- White-board (one participant at front of the room talks and makes notes on the white board)

In each meeting there were four participants at six possible

1424403677/06/$20.00 ©2006 IEEE

Meeting Event	Train	Test
Discussion	48	49
Monologue 1	14	12
Monologue 2	10	13
Monologue 3	10	14
Monologue 4	9	10
Note-taking	6	3
Presentation	11	18
White-board	16	20
Total	124	139

Table 1. Number of meeting events in different sets

positions: four seats plus whiteboard and presentation board. The number of different meeting events in the different data sets is summarized in table 1.

The database comprises a total of 59 scripted meetings with two disjoint sets of participants. A fixed training set makes use of 30 videos, while the remaining 29 are used throughout evaluation.

3. MULTI-MODAL FEATURE EXTRACTION

For each participant person-specific features were extracted from the cameras, the lapel microphones, and the microphone array. Therefore we make use of visual as well as audio features. The person-specific video features are:

- head vertical centroid
- head eccentricity
- right hand horizontal centroid
- right hand angle
- right hand eccentricity
- head and hand motion

For each video frame areas of skin color are detected by a Gaussian mixture model. Next the greatest skin color blobs are identified as face using a face detector and described by the vertical centroid and eccentricity. From the remaining blobs the one with the rightmost horizontal position is regarded as hand and is represented by its horizontal position, eccentricity, and angle. For more detail on the video features please refer to [8].

In addition to the visual features we also used person-specific audio features extracted from the lapel microphones and the microphone-array:

- speech activity from each seat
- speech relative pitch
- speech energy
- speech rate

As speech activity measure SRP-PHAT was used. Pitch was extracted using a SIFT algorithm and normalized to the mean value. All used features and methods to derive them are explained in more detail in [8].

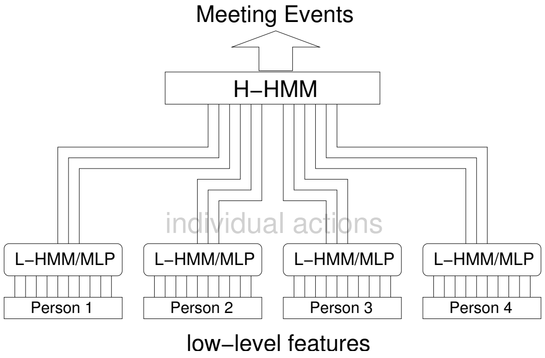

Fig. 1. Overview of the two-layered system

In addition to the individual multi-modal features group features were extracted from the white-board and projector-screen area. In detail they include the speech activity from the white-board and projector screen as audio features. From the visual information the mean difference between a current frame and a reference background image is used.

4. SYSTEM OVERVIEW

We propose a two stage system for the segmentation and recognition of meetings into meeting events as shown in figure 1. In the lower level, individual actions are recognized to build a bridge between the low-level features and the meeting events. We distinguish a small set of $N_I = 3$ individual actions. These consist in:

- *speaking* (one participant is speaking)
- *writing* (one participant is taking notes)
- *idle* (one participant is neither speaking nor writing)

A larger set of individual actions was used alternatively, but as the results later will confirm this brought no improvement. The low-level feature-frames are classified by continuous Gaussian mixture HMM and a multilayer perceptron network respectively. The results are subsequently fed forward as posteriors to the next layer which consists of continuous Gaussian mixture HMM to provide a segmentation and recognition of meeting events via the Viterbi algorithm.

4.1. Individual Action Recognition

Two different approaches to recognize individual actions were tested. The first model we applied was a standard HMM with various numbers of states and gaussians to which we refer a "L-HMM" in the following as it models the lower layer of our system. Linking to the higher level is accomplished by producing a likelihood of each frame of each individual action. This is done as follows: Let $a^t = (a_1^t, \ldots, a_{N_I}^t) \in \mathbb{R}^{N_I}$ denote a vector of dimension equal to the number of individual actions. The a_i^t indicate the likelihood of the model i at time t

for an observation sequence $\boldsymbol{\xi}_{1:t} = \xi_1, \xi_2, \ldots, \xi_t$. During decoding the forward variable $\alpha(i,t) = P(\boldsymbol{\xi}_{1:t}, q_t = i)$ defines the likelihood of a model having generated the sequence $\boldsymbol{\xi}_{1:t}$ and being in state i at time t.

$$P(\omega_i|\boldsymbol{\xi}_{1:t}) = \frac{\sum_i^{N_I} \alpha(i,t)}{\sum_j^{N_{tot}^s} \alpha(j,t)} \quad (1)$$

Adding up the α_i over all classes N_I and normalizing over all states N_{tot}^s as shown in eq. 1 produces a probability for each class ω_i where the probabilities over all classes sum to one. These probabilities are then fed forward to higher level HMM (denoted as "H-HMM" in the following) to recognize the meeting events. To avoid numbers beyond the computational accuracy while multiplying probabilities during decoding a segmentation of the audio-visual features is done via Viterbi-decoding. Next for each detected segment the probabilities for each frame are calculated.

The second approach to recognize individual actions comprises a multi-layer perceptron artificial neural network (MLP). Here we use a straight forward layout with one hidden layer. The N_I outputs of each frame, q_i, $i = 1, \ldots, N_I$, of the MLP are normalized using the softmax function

$$p_i = \frac{e^{q_i}}{\sum_{j=1}^{N_I} e^{q_j}} \quad (2)$$

Likewise the output can be interpreted as probabilities for each class. These are then passed on to higher level HMM to recognize meeting events.

4.2. Meeting Event Recognition

The results from the lower level individual action recognizers from each participant are concatenated together with the group-level features into a $(N_I \times N_P + N_{GP})$-dimensional vector (where N_P is the number of participants and N_{GP} is the dimension of the group features). These so obtained vectors are subsequently fed forward to continuous Gaussian mixture HMM that provide a segmentation and recognition of the meeting events via the Viterbi algorithm.

5. EXPERIMENTS

Prior to presentation of results we describe measures used throughout evaluation. Results thereof are provided for two different meeting event classification approaches: Two-layered HMM and the hybrid MLP-HMM approach.

5.1. Performance measures

We use the *frame error rate* (FER) and the *accuracy* as measures to evaluate the results of the meeting event recognition. The FER is used to evaluate results of individual action classifiers. It is defined as one minus the ratio between the number of correctly recognized frames and the number of total frames: $FER = (1 - \frac{correctframes}{totalframes}) \times 100\%$. For the evaluation of the segmentation performance we use the commonly accepted accuracy measurement defined as one minus the sum of insertions (Ins), deletions (Del), and substitutions (Sub), divided by the total number of events in the ground truth defined by manually labeling the meeting corpus: $Accuracy = (1 - \frac{Subs+Del+Ins}{TotalEvents}) \times 100\%$.

Individual Action	**Train**	**Test**
Idle	1423	1485
Speaking	1057	1022
Writing	351	476
Total	2831	2983

Table 2. Number of individual actions in different sets

Model	**FER in %**
L-HMM (4 States, 10 Gaussians)	9.63
MLP	10.17

Table 3. Results of the individual action recognition

5.2. Individual Action Recognition

In extensive tests with a wide variety in the number of states and Gaussians the optimal configuration for the task of recognizing individual actions was searched for. Results are presented in Table 3 for the low-level HMM and MLP respectively. As can be seen the frame error rate is around 10% in both models. The confusion matrix for L-HMM is shown in Table 4. In general all three individual actions are well recognized. *Writing* tends to be confused with *speaking* but only in a low number of occurrences. Amazingly idle is detected rather well even though it covers all other possible activities (e.g. pointing, nodding), in contrary to the two other well-defined actions. Although the deletion rate is rather high this is of less importance as we only use the segmentation as a rough grid to avoid computational accuracy problems. Tests with a greater number of individual actions were discarded as the FER clearly revealed that these larger number of actions cannot be modeled by use of the current systems and features.

in %	**Idle**	**Speaking**	**Writing**	DEL
Idle	97.38	0.33	2.30	38.25
Speaking	1.46	91.16	7.55	46.87
Writing	0.30	2.67	97.03	29.20
INS	1.08	0.0	3.16	

Table 4. Confusion matrix of Viterbi-decoded individual actions using HMM with insertions (INS) and deletions (DEL)

in %	Idle	Speaking	Writing
Idle	92.71	4.06	3.24
Speaking	15.16	84.39	0.45
Writing	14.81	0.46	84.73

Table 5. Confusion matrix of frame-wise recognized individual actions using MLP

Method	Accuracy in%
Single-layered HMM	77.57
Two-layered HMM	87.77
MLP-HMM	90.65

Table 6. Results of meeting event recognition

A similar situation is given by the results of a MLP as shown in Table 5. There again *idle* is detected quite well, whereas *speaking* and *writing* tends to get confused with *idle*. But there is a strong distinction between the two latter, as the other is never falsely detected in more than 0.46% of all cases.

5.3. Meeting Event Recognition

For recognizing meeting events we use the outputs of the L-HMM and MLP respectively. These are concatenated as described in section 4.2. Next a H-HMM is trained on this data. For comparison we also use a single-layer HMM using the low-level features described in section 3 as input. The results are presented in Table 6 where the accuracy for each model is shown separately. As can be clearly seen, the two-layered approach significantly outperforms the conventional Gaussian-Mixture HMM. There is a gain in the accuracy of about ten percent absolute. Furthermore our hybrid MLP-HMM approach exceeds the latter by another 2.88% absolute. This shows the discriminative power of our suggested procedure. The confusion matrix, which is almost a diagonal matrix and therefore not shown here, reveals the performance in more detail. *Discussion* is only confused with *monologue 1* and *monologue 2* once, whereas *whiteboard* is twice taken as *presentation*. There are also only three meeting events inserted (*discussion*, *monologue 4*, and *white-board* once each), and altogether six events deleted from the overall sum of 139 meeting events.

6. SUMMARY AND CONCLUSION

In this work we presented an approach for the automatic segmentation of recorded meetings into meeting events. Combining artificial neural nets and HMM results in a highly discriminative system. Conventional models were clearly outperformed by our suggested two-stage approach. Even a two-layered HMM framework is surpassed. The accuracy for discriminating complex meeting events is in similar high compared to other research groups although a direct comparison is not possible due to a slightly different labeling of the meeting events. However tests using the same labels as suggested in [8] are scheduled.

In our future research we plan to incorporate other types of artificial neural nets in the lower layer as well as Dynamic Bayesian Networks in the higher level for their ability of representing complex stochastic processes.

7. ACKNOWLEDGMENTS

This work was partly supported by the EU 6th FWP IST Integrated Project AMI (FP6-506811, publication AMI-152). We also thank Dong Zhang (IDIAP) for providing the features and part of the ground-truth labels.

8. REFERENCES

[1] Nelson Morgan, Don Baron, Jane Edwards, Dan Ellis, David Gelbart, Adam Janin, Thilo Pfau, Elizabeth Shriberg, and Andreas Stolcke, "The meeting project at ICSI," in *Proceedings of the Human Language Technology Conference*, San Diego, CA, March 2001.

[2] Klaus Zechner, "Automatic generation of concise summaries of spoken dialogues in unrestricted domains," in *Proceedings of the 24th ACM-SIGIR International Conference on Research and Development in Information Retrieval*, New Orleans, LA, September 2001.

[3] Michael Bett, Ralph Gross, Hua Yu, Xiaojin Zhu, Yue Pan, Jie Yang, and Alex Waibel, "Multimodal meeting tracker," in *Proceedings of RIAO2000*, Paris, France, April 2000.

[4] Rainer Stiefelhagen, "Tracking focus of attention in meetings," in *IEEE International Conference on Multimodal Interfaces*, Pittsburgh, PA, USA, October 14–16 2002.

[5] I. McCowan, S. Bengio, D. Gatica-Perez, G. Lathoud, F. Monay, D. Moore, P. Wellner, and H. Bourlard, "Modeling human interaction in meetings," in *Proceedings of International Conference on Acoustics, Speech and Signal Processing*, Hong Kong, April 2003.

[6] Stephan Reiter and Gerhard Rigoll, "Segmentation and classification of meeting events using multiple classifier fusion and dynamic programming," in *IEEE Proceedings of the International Conference on Pattern Recognition (ICPR)*. IEEE Computer Society, August 2004, pp. 434–437.

[7] Stephan Reiter and Gerhard Rigoll, "Multimodal meeting analysis by segmentation and classification of meeting events based on a higher level semantic approach," in *Proceedings of the 30th International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Philadelphia, USA, March 2005.

[8] Dong Zhang, Daniel Gatica-Perez, Samy Begio, Iain McCowan, and Guillaume Lathoud, "Modeling individual and group actions in meetings: a two-layer hmm framework," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition, Workshop on Event Mining in Video (CVPR-EVENT)*, Washington DC, July 2004.