DETECTING GROUP INTEREST-LEVEL IN MEETINGS

Daniel Gatica-Perez, Iain McCowan, Dong Zhang, and Samy Bengio

IDIAP Research Institute
Rue de Simplon 4, CH-1920 Martigny, Switzerland
{gatica, mccowan, zhang, bengio}@idiap.ch

ABSTRACT

Finding relevant segments in meeting recordings is important for summarization, browsing, and retrieval purposes. In this paper, we define relevance as the interest-level that meeting participants manifest as a group during the course of their interaction (as perceived by an external observer), and investigate the automatic detection of segments of high-interest from audio-visual cues. This is motivated by the assumption that there is a relationship between segments of interest to participants, and those of interest to the end user, e.g. of a meeting browser. We first address the problem of human annotation of group interest-level. On a 50-meeting corpus, recorded in a room equipped with multiple cameras and microphones, we found that the annotations generated by multiple people exhibit a good degree of consistency, providing a stable ground-truth for automatic methods. For the automatic detection of high-interest segments, we investigate a methodology based on Hidden Markov Models (HMMs) and a number of audio and visual features. Single- and multi-stream approaches were studied. Using precision and recall as performance measures, the results suggest that the automatic detection of group interest-level is promising, and that while audio in general constitutes the predominant modality in meetings, the use of a multi-modal approach is beneficial.

1. INTRODUCTION

The development of methods to segment and extract *relevant* segments from a collection of meeting recordings is important for a number of information retrieval (IR) tasks. For browsing, summarization, and retrieval, in addition to the use of speech transcription-based IR techniques, which aim to account for what is *said* in a meeting, the automatic recognition of what is *acted* by the meeting participants represents a valuable feature.

Obviously, what constitutes a relevant meeting segment requires further clarification. Motivated by concepts in social psychology, which highlight the group and multimodal nature of communication, recent work has viewed meetings as sequences of non-overlapping multimodal actions performed by the group of participants (e.g. turn-taking), thus implying that such actions are relevant to segment and recognize [8, 4]. In this paper, we use the concept of *group interest-level* to define relevance, phrasing it as the degree of engagement that meeting participants display as a group during their interaction, as perceived through the audio and visual modalities by an external observer. With this definition, relevant segments become those that depict high group interest-level. From an application standpoint, one could expect that such segments would also be likely to interest other observers, e.g. using a meeting browser.

The modeling of interest level or other closely related concepts has been recently explored in multi-person conversational settings [10, 6, 5]. Using only the speech modality, and defining speech utterances as the basic units, the work in [10] defined the concept of *hot-spots* in much the same way we define high interest level, relating it to the concept of activation in emotion modeling [3]. Furthermore, the work in [6] defined *emphasis* for speech utterances, acknowledging that this concept and emotional involvement might be acoustically and perceptually similar. The work in [5] described a PDA-based system in which meeting participants manually input their interest level, which is incorporated into a conversation analysis module to extract high-interest segments. Interest level has also been explored in computer-assisted learning [9].

We address the problem using low-level audio-visual features and statistical models. Given a stable ground-truth for supervised learning, the HMM-based continuous recognition approach jointly generates a meeting segmentation and the classification of the meeting segments. We experimented with a number of features and schemes for data fusion, and obtained enrouraging results through the use of standard features and models. We are not aware of any previous studies on the use of multiple modalities for the recognition of group-based interest level in meetings as defined here.

The paper is organized as follows. Section 2 presents and discusses the group interest-level annotation process. Section 3 presents our approach in detail, including a description of the models used for recognition and the audio-visual features. Section 4 presents the experimental results and discusses our findings. Section 5 concludes the paper.

2. ANNOTATING GROUP INTEREST-LEVEL

2.1. Meeting corpus

The public corpus we use is a subset of the one first presented in [8], and consists of 50 five-minute, four-participant meetings, recorded in a room equipped with three cameras and 12 microphones. Although the meetings were recorded according to a script for turn-taking patterns, the participants' behavior was unconstrained and reasonably natural in terms of emotional engagement.

2.2. Protocol and discussion

The corpus was first annotated for the perceived group interest-level. Unlike events for which a definite ground-truth exists, e.g. speech words or physical attributes, it is not trivial to consistently annotate subjective phenomena such as interest-level, and the annotation approach should depend on the particular task at hand ([1] contains an overview of different approaches).

This work was carried out in the framework of the Swiss National Center of Competence in Research on Interactive Multimodal Information Management (IM2), and the European projects MultiModal Meeting Manager(M4), and Augmented Multi-party Interaction (AMI, pub. AMI-41). The first three authors contributed equally to this work.

In contrast to [10], in which hot-spot annotation was based on assigning labels to utterance units, we employed an interval coding scheme [1]. As such, our work does not assume that high interest-level periods align with utterance boundaries, and no prior segmentation of the data is required. While viewing an audio-visual recording of a meeting, annotators were asked to rate the group interest-level on a discrete scale of 1-5 (very low to very high, 3 being neutral), for every 15 second interval. As a guide, the group interest-level was described as the perceived degree of interest or involvement of the majority of the group, and examples of activity indicating interest were given, including note-taking, focused gaze, and avid participation in discussion. Each meeting was presented to the annotators as a synchronized composition of the three camera views and one audio signal, available at `mmm.idiap.ch`.

While interval coding has limitations, it was chosen for practical considerations: notably its speed and simplicity. Interval coding allows annotators to code a meeting in one-pass, and in roughly real-time. It also alleviates the need to define precise onset times for each 'event', which was felt to be unrealistic and unnecessary for the phenomenon studied here. A 15 second interval duration was chosen, as it is shorter than the expected duration of events (e.g. a period of high interest) but still contains sufficient evidence on which to base a rating. Using a numeric scale, rather than a set of categorical labels, also holds a number of practical advantages: it encodes the natural relationship between the different labels (e.g. 4 is 'closer' to 5 than to 2), and it facilitates analysis and combination of multiple annotations. We note that, while we were not interested in distinguishing 5 grades of interest-level in experiments, annotators preferred to have such a range to grade (e.g. to distinguish 'interested' from both 'neutral' and 'highly interested'), and this also serves to improve consistency when scores are merged to a smaller number of categories.

Each meeting was annotated by two independent annotators, taken from a pool of 12 people. The raw annotations were processed to normalize for annotator bias, analyze inter-annotator agreement, and then combine ratings across the two annotators. Annotator bias was compensated by normalizing the scores across all meetings for a single annotator to have zero mean and unity variance. Due to the ordinal and interval nature of the annotations, inter-annotator agreement was simply assessed by calculating the correlation between the two sequences of scores, yielding a correlation coefficient of 0.68 over the corpus. The mean of the two scores was taken as the ground-truth score for each interval.

For the current experiments, we are only interested in distinguishing periods of high-level interest from all others. For this purpose, the processed scores were used to generate the required binary labels: approximately the top 20% of interval scores were labelled as high-interest, while the rest were labelled as neutral. After annotation, the number of frames for high and neutral interest-level for training and testing sets appear in Table 1. The final ground-truth segmentation was defined by grouping contiguous intervals having the same label.

Following the annotation process, we asked annotators to provide a list of the informal rules they applied in distinguishing high interest periods. The responses corresponded in general to the provided guidelines, but also included phenomena such as discussion segments where most people contribute within a short interval, laughter, and interaction of others with the main speaker through head gestures or audio back-channels ('yeah', 'uh-huh', etc). We note that the above list is multimodal in nature, and most phenomena could be reasonably observed using automatic techniques.

Interest level	**Train**	**Test**
high	8672	7017
neutral	35847	22433
total	44519	29450

Table 1: Number of samples for interest level in different data sets.

3. OUR APPROACH

In this section, we describe our approach for continuous recognition of high (and neutral) group interest-level from audio-visual data. We first describe the specific statistical models, and then present the extracted features.

3.1. Statistical models

We investigated two classic HMM recognition strategies that, similar to the approach proposed in [8], produce both a segmentation of a meeting sequence and the classification of each of the segments. The first one is the basic early integration approach, where all desired streams (audio, visual, or audio-visual) are aligned, synchronized, and concatenated to form the input observation vector. The second model is a multi-stream HMM (MS-HMM), which was only used for audio-visual fusion. In this case, the audio and visual streams are trained independently, and the outputs of both modalities are merged at the state level during decoding, by a convex combination of the outputs, defined by a weight parameter (ω).

3.2. Audio-visual features

It is expected that some features from the recent literature could be appropriate for our task [10, 6]. Additionally, from the annotation guidelines and the annotators' feedback, visual detectors of increased activity, note-taking, and gaze, and audio detectors of laughter could be useful too. As a first step, instead of dealing with more targeted features, in this paper we extracted a set of generic features, including audio features derived from microphone arrays and lapel microphones, and visual features extracted from skin color blobs from each participant [8]. This initial audio-visual set was later used in a feature selection procedure (see section 4).

For audio features, a speech activity measure (SRP-PHAT) was first estimated at four seated locations from the microphone array signals. Energy, pitch, and speaking rate were then estimated from each lapel. Following segmentation according to [7], these features were computed only on speech segments, zeroing all silence segments. We used the SIFT algorithm for pitch, and a combination of estimators for speaking rate [8].

For video features, skin-color head and right-hand blobs for each participant were first extracted using a standard approach [8]. A number of features were then computed, including global person motion (the addition of head+hand motion), and features related to person pose (eccentricity and orientation for hand blobs, and a rough measure of head orientation).

4. EXPERIMENTS AND RESULTS

This section describes the experiments to recognize high interest level in meetings. We first describe the measures used to evaluate our results. We then describe the feature selection process. Finally we present results and discuss our findings.

4.1. Performance measures

For our two-class classification problem, the Action Error Rate (identical to the Word Error Rate in ASR), as used previously in [8], is no longer appropriate as an evaluation measure. Instead, the performance was measured in terms of precision (pr) and recall (rc), using frames as the basic unit. If N_c, N_f and N_d denote the

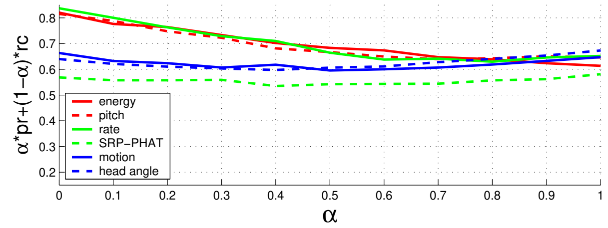

Figure 1: EPCs for individual audio and video feature selection.

number of high-level frames correctly detected, falsely accepted, and falsely rejected by the system, respectively, the measures are defined as $pr = N_c/(N_c+N_f)$ and $rc = N_c/(N_c+N_d)$. Instead of computing the typical precision-recall curves, we opted for the Expected Performance Curve (EPC) proposed in [2], which has shown to provide a fairer comparison between models. The procedure optimizes a convex combination of the individual performance measures, $ep = \alpha * pr + (1-\alpha) * rc$, on the validation set during the training procedure used for parameter selection, and then plots pr, rc, or ep on the test set as a function of the convex parameter.

4.2. Experimental setup

The public corpus is divided into 30 meetings for training, and 20 for testing. For training, we used a six-fold cross-validation procedure to select the best HMM parameters, splitting the training set into training and validation subsets. After the best model parameters were chosen, we re-trained models on the whole training set and use the parameters on the test set. The procedure was repeated for 10 different α values, in steps of 0.1. The number of states per class, and number of Gaussians per mixture could range between 1 and 20 in both cases. For MS-HMM, 10 weight combinations were used, also in steps of 0.1.

4.3. Feature selection

To select a subset of features from the complete set described in Section 3.2, we used a simple empirical method, employing a basic HMM on individual features, and choosing the best ones for further combination (see Section 4.4.) From Fig. 1, we selected the three best audio features: speech energy, speaking rate, and speech pitch. Interestingly, these results are in accordance with recent literature that mentions pitch as a prosodic cue that shows correlation with individual level of involvement [10], and as a feature used for recognition of emphasis [6], both on speech utterances. Note however that the results here are defined over segments rather than over utterances, and assigned to a group of meeting participants rather than to individuals. Also note that the SRP-PHAT features did not perform as well by themselves, although they are implicitly used in the speech/silence segmentation used to define the prosodic features. Regarding video, the two best features that were selected were person motion and head angle. Overall, this results in three audio and two video features per participant, which in principle allows the HMM to model correlation between participants.

Additionally, to explore the potential benefits of feature fusion at the group level, we reduced the feature space by averaging each of the features over all the participants. This results in a single vector of three audio and two video features.

4.4. Studied cases

We investigated the following feature-model combinations:

1. *HMM, audio-only*, with the 3 best (x4) features.
2. *HMM, video-only*, with the 2 best (x4) features.
3. *HMM, audio-video*, with the above 5 (x4) features.
4. *MS-HMM, audio-video*.

Regarding feature fusion at the group level, we studied three more cases:

5. *HMM, audio-only, 3 group-fused features*.
6. *HMM, audio-video, 5 group-fused features*.
7. *MS-HMM, audio-video, 5 group-fused features*.

4.5. Results and discussion

In the following, the EPCs show the performance of the algorithm for a number of precision-recall combinations. When α is close to 0, recall dominates; precision does so when α approaches 1. As a simple baseline, if all frames are labeled as high interest level, given that roughly 20% of the frames in the ground-truth, the results are $rc = 1.0$, $pr = 0.2$, and ep varies between 1.0 and 0.2 as α varies from 0 to 1.

The results can be summarized as follows.

Single modalities. The EPCs for cases 1 and 2 are shown in Figs. 2(a) and 2(b), respectively. Comparing to Fig. 1, the audio feature combination is clearly better than the individual audio features to recognize high interest-level. Audio-only is also better than video-only. Note that when recall is the dominant measure in ep (small α), audio-only recall is usually high ($rc \geq 0.8$) at a not-so-low precision ($pr \geq 0.55$). On the contrary, when precision is the dominant measure (large α), audio-only precision is good ($pr \geq 0.7$), but recall degrades significantly ($rc \leq 0.35$). Visual features are "noisier", in that the overall ep is lower, particularly for small α, but in general they show a more consistent trend (especially with respect to precision) across different values of α. We note that when precision is dominant ($\alpha \geq 0.8$), the performance is essentially the same for both modalities. These somewhat complementary roles raised some expectations for modality fusion.

Audio-visual fusion. The EPC for case 3 is shown in Fig. 2(c). A comparison of ep for all methods is shown in Fig. 2(e). The ep measure for $\alpha \in [0, 0.6]$ has degraded w.r.t. audio-only, due to a reduction in recall. This can be explained by the ambiguity introduced by the visual features.In contrast, an improvement was observed in the range [0.6,1], where video features contribute to increase both recall and precision. The fact that a-v performed better than audio-only for a range of α in an interesting result in itself, and also motivated the use of the MS approach.

The EPC for case 4, for the best weight combination (0.9 on audio, 0.1 on video), is shown in Fig. 2(d). Given the predominance given to audio, the ep measure has degraded gracefully w.r.t. audio-only for $\alpha \in [0, 0.55]$, but has maintained the improvement for all other cases, thus outperforming the basic HMM, and confirming that modality combination is beneficial if carefully done.

Feature fusion at the group level. Fusing features at the group level might be advantageous both for dimensionality reduction and for exploring schemes for combinations of individual responses. We found that feature averaging essentially produced improvements in every case (5, 6, and 7) w.r.t. to the original ones (1, 3, and 4, respectively). To compare these six cases, the EPCs are shown in Fig. 2(f). In brief, the best performance was obtained for audio-only, group-fused features for $\alpha \in [0, 0.55]$, while the MS-HMM with group-fused features, and a weight combination of (0.7,0.3) on audio-video, performed the best for $\alpha \in [0.55, 1]$. We speculate that dimensionality reduction plays an important role in the observed improvements. As a summary, precision/recall values for three typical cases, $\alpha = 0, 0.5$, and 1, are shown in Table

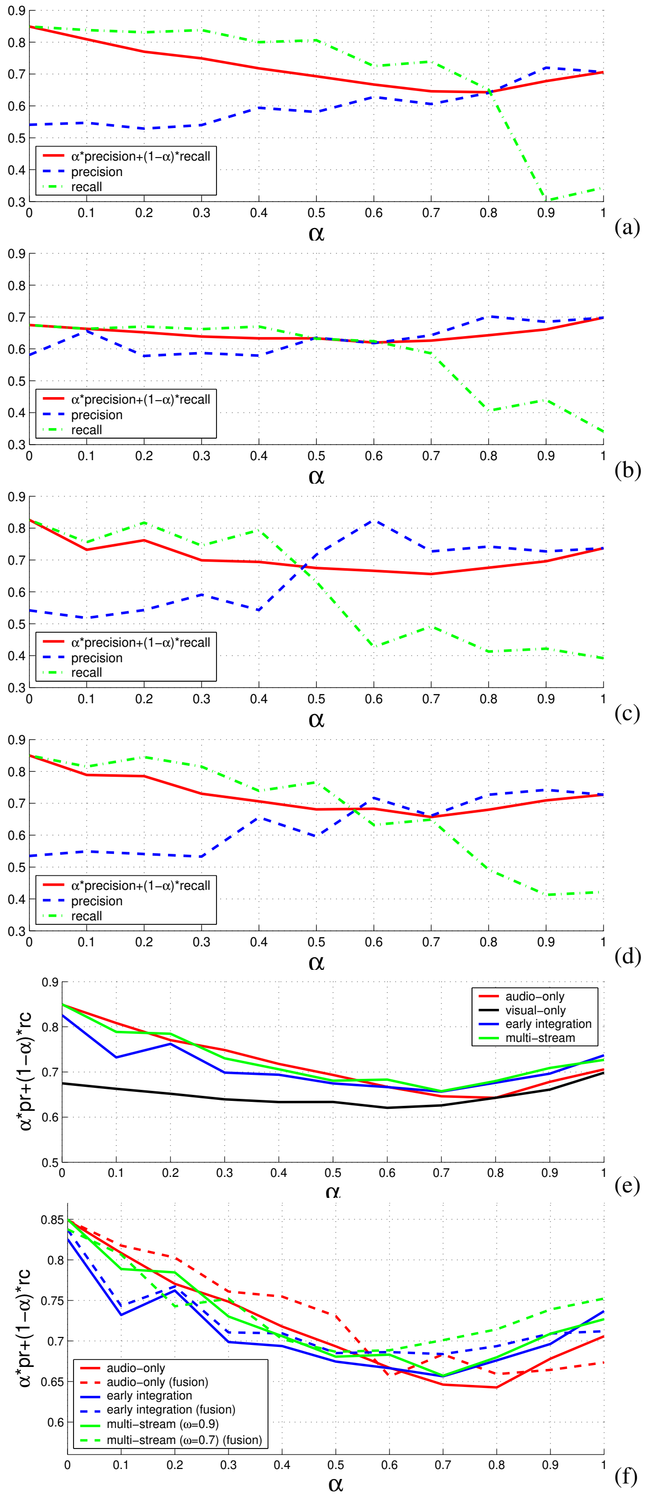

Figure 2: EPCs for (a) audio-only; (b) visual-only; (c) a-v HMM; (d) a-v, MS-HMM, (e) joint display; (f) feature fusion at the group level. EPCs for group-fused cases are shown in dashed line; original cases in continuous line.

case	$\alpha = 0$		$\alpha = 0.5$		$\alpha = 1$	
	pr	rc	pr	rc	pr	rc
1	0.54	0.85	0.58	0.80	0.70	0.34
5	0.63	0.85	0.63	0.84	0.67	0.54
3	0.54	0.83	0.72	0.63	0.74	0.39
6	0.56	0.84	0.63	0.74	0.71	0.60
4	0.54	0.85	0.60	0.77	0.73	0.42
7	0.59	0.84	0.77	0.60	0.75	0.55

Table 2: Precision/recall values for three values of α.

2. The investigation of other ways of merging features at the group level will be the subject of further study.

5. CONCLUSION

We investigated the viability of recognizing segments of high interest-level in meetings, using low-level audio-visual features, and statistical sequence models. While preliminary, our results are encouraging, and seem to agree with recent work conducted on the audio modality. Furthermore, our work provided an initial result on the issue of modality combination for performance improvement. Future work includes the annotation of group interest level as a function of the activation level of the meeting participants, as defined in the emotion literature [3], the investigation of higher-level features appropriate for the task, and the evaluation of our methodology on fully real (rather than scripted) meetings.

Acknowlegments. We thank Guillaume Lathoud for his work on audio feature extraction.

6. REFERENCES

[1] R. Bakeman and J.M. Gottman. *Observing Interaction : An Introduction to Sequential Analysis*. Cambridge UP, 1997.

[2] S. Bengio and J. Mariéthoz. The expected performance curve: a new assessment measure for person authentication. In *Proc. Odyssey*, Toledo, May 2004.

[3] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. Taylor. Emotion recognition in HCI. *IEEE Signal Processing Magazine*, Jan. 2001.

[4] A. Dielmann and S. Renals. Dynamic Bayesian networks for meeting structuring. In *Proc. ICASSP*, Montreal, May 2004.

[5] N. Eagle and A. Pentland. Social network computing. In *Proc. UBICOMP*, Seattle, Oct. 2003.

[6] L. Kennedy and D. Ellis. Pitch-based emphasis detection for characterization of meeting recordings. In *Proc. ASRU*, St Thomas, Dec. 2003.

[7] G. Lathoud, I. McCowan, and D. Moore. Segmenting multiple concurrent speakers using microphone arrays. In *Proc. Eurospeech*, Geneva, Sep. 2003.

[8] I. McCowan, S. Bengio, D. Gatica-Perez, G. Lathoud, F. Monay, D. Moore, P. Wellner, and H. Bourlard. Modeling human interactions in meetings. In *Proc. ICASSP*, Hong-Kong, Apr. 2003.

[9] S. Mota and R. Picard. Automated posture analysis for detecting learner's interest level. In *Proc. CVPR Workshop on CVPR for HCI*, Madison, Jun. 2003.

[10] B. Wrede and E. Shriberg. Spotting hotspots in meetings: Human judgments and prosodic cues. In *Proc. Eurospeech*, Geneva, Sep. 2003.