Automatic Detection of Interaction Groups

Oliver Brdiczka, Jérôme Maisonnasse, Patrick Reignier

Laboratoire GRAVIR
INRIA Rhône-Alpes
655 Av. de l'Europe
38330 Montbonnot, France.

{brdiczka, maisonnasse, reignier}@inrialpes.fr

ABSTRACT

This paper addresses the problem of detecting interaction groups in an intelligent environment. To understand human activity, we need to identify human actors and their interpersonal links. An interaction group can be seen as basic entity, within which individuals collaborate in order to achieve a common goal. In this regard, the dynamic change of interaction group configuration, i.e. the split and merge of interaction groups, can be seen as indicator of new activities. Our approach takes speech activity detection of individuals forming interaction groups as input. A classical HMM-based approach learning different HMM for the different group configurations did not produce promising results. We propose an approach for detecting interaction group configurations based on the assumption that conversational turn taking is synchronized inside groups. The proposed detector is based on one HMM constructed upon conversational hypotheses. The approach shows good results and thus confirms our conversational hypotheses.

Categories and Subject Descriptors

I.2.10 **[ARTIFICIAL INTELLIGENCE]**: Vision and Scene Understanding - *Perceptual reasoning.*

General Terms

Algorithms, Human Factors, Experimentation, Performance.

Keywords

Clustering interaction groups, Hidden Markov Model, speech detection, conversational analysis, intelligent environment, ubiquitous computing.

1. INTRODUCTION

Ubiquitous computing enables computer systems to sense and to respond to human activity. Human actors need to be identified in order to perceive correctly their activity. In intelligent environments more and more devices are capable of perceiving user activity and offering services to the user. Offering services means to supply a system reaction or an interaction at the most appropriate moment, aligned with the activity of the users. Addressing the right user at the correct moment is essential. Thus we need to detect potential users and their connection while doing an activity.

The identification of the current group configuration of the users is necessary to analyze activity. In a physical environment, several individuals can form one group working on the same task, or they can split into subgroups doing independent tasks in parallel. The dynamics of group configuration, i.e. the split and merge of small interaction groups, allows us to perceive the appearing of new activities. We assume that a change in group configuration is strongly linked to a change in activity, at least to an interruption of the current activity. The fusion of several independent small groups is seen as important information for detecting a change of the current activity, on a local or global level. For example, people attending a seminar tend to form small groups discussing different topics before the seminar starts. When the lecturer arrives, these small groups merge and form a big group listing to the lecture. In this example, the fusion of several small groups to one big group can be used to detect the beginning of a seminar. In the same manner, the split of the big group into several small groups can indicate a pause or the end of the lecture. The change in group configuration is thus a strong indicator of new activities as well as of activities that are linked to a particular group configuration (for example a lecture).

In this paper, we propose a method for the dynamic detection of small group configurations based on Hidden Markov Models. The method relies on speech activity detection as sensor information for interacting individuals. We focus thus on verbal interaction, which further implies a minimum size of two individuals for one group (assuming that isolated individuals do not speak). The method has been tested in experiments recording meetings of 4 individuals.

2. PREVIOUS AND RELATED WORK

The recognition of human activity based on speech events is often used in the context of group analysis. In general, the group and its members are defined in advance. The objective is then to use frequency and duration of speech contributions to recognize particular key actions executed by group members [11] or to analyse the type of meeting in a global manner [3]. However, the detection of dependencies between individuals and their membership in one or several groups is not considered. Analysing

ICMI'05, October 4–6, 2005, Trento, Italy.

large amounts of data from recordings of interactions enables the reconstruction of social networks for a number of individuals [4]. The detection and analysis of conversations is then necessary. The automatic detection of conversations using mutual information [1], in order to determine who speaks and when, needs an important duration of each conversation.

In this paper, we want to develop a real-time detector for interaction groups. This detector should be robust and as general as possible. The objective is to define the limits combining several individuals for doing the same intended activity. This activity is in most cases the main activity [8], provided that verbal interaction needs a certain level of attention. Note that the activity of a group of individuals can, in particular moments, attract the attention of other individuals, which means a short-term collective focusing on one activity [10]. Goffman calls this focusing "participation framework" [9]. Our objective is to visualize this dynamics in order to enable intelligent environments to use this information for the recognition of activity.

Conversational analysis [13] and social psychology pointed out a certain number of important points concerning interpersonal interaction. Verbal interactions within a group are regulated [7], intentional and composed of successive conversations aiming at acting on a common ground [5] [6]. These statements will allow us to formulate conversational hypotheses as significant criteria for the detection of interaction groups.

3. APPROACH

We present an approach based on Hidden Markov Models [12]. A Hidden Markov Model is a stochastic process where the evolution is managed by states. The series of states constitute a Markov chain which is not directly observable. This chain is "hidden". Each state of the model generates an observation. Only the observations are visible. The objective is to derive the state sequence and its probability, given a particular sequence of observations. Hidden Markov Models have been used with success in speech recognition [14], sign language recognition [15], hand-writing gesture recognition [16] and many other domains. We use a HMM approach due to the dynamics of group split and merge as well as the noisy character of speech activity detection data.

3.1 Speech Activity Detection

The observations used as input for the HMM are generated from speech activity data. An automatic speech detector [17] parses multi-channel audio input and detects which individual stops and starts speaking. The observations of the Hidden Markov Model are a discretization of speech activity events sent by this detector. One observation is a vector containing a binary value (speaking, not speaking) for each individual that is recorded. This vector is transformed to a 1-dimensional discrete code used as input for the HMM (see Table 1). The automatic speech detector has a sampling rate of 62.5 Hz, which corresponds to the generation of an observation vector every 16 milliseconds.

Table 1. Observations of a Hidden Markov Model for a meeting of 4 individuals

Observation Number	Speech Activity			
	A	B	C	D
0	0	0	0	0
1	0	0	0	1
2	0	0	1	0
3	0	0	1	1
4	0	1	0	0
5	0	1	0	1
6	0	1	1	0
7	0	1	1	1
8	1	0	0	0
9	1	0	0	1
10	1	0	1	0
11	1	0	1	1
12	1	1	0	0
13	1	1	0	1
14	1	1	1	0
15	1	1	1	1

3.2 Classification Using HMMs

A first approach is to classify different group configurations using several HMMs. Each a priori group configuration class is associated with a Hidden Markov Model. The classification system is composed of *n* HMMs, where *n* corresponds to the number of possible group configurations for the recorded individuals. During a training stage, HMM parameters are estimated from a data set. This set is composed of several example concurrencies for each group configuration class. To classify, the probability that an unknown observation sequence was produced by the Hidden Markov Model is calculated using the forward-backward procedure [12]. The HMM with the highest probability for the given observation sequence determines the current group configuration.

The number of states of the HMMs for the different group configurations is unknown a priori and needs to be fixed before training stage. K-means algorithm [12] is used for an initial training of the parameters of the HMMs, while Baum-Welch re-estimation formulas [12] are used for further training. Both algorithms are run with a fixed number of states for the HMM to train. To determine the optimal number of states for the HMMs, we varied the number of fixed states for the training. We evaluated the HMMs using audio recordings of two meetings of four individuals. Four HMMs for the four possible group configurations have been trained with different numbers of states. We did a cross-validation by training HMMs with the group configurations of the first meeting and classifying the group configurations of the second meeting and vice versa. Table 2 and 3 show the results.

Table 2. Correct classification of group configurations of Meeting 1 and Meeting 2 with HMMs trained on group configurations of Meeting 1 (training set: Meeting 1, test set: Meeting 2).

States	*1*	*2*	*4*	*8*
Meeting 1	0.72	0.29	0.27	0.29
Meeting 2	0.51	0.23	0.23	0.19

Table 3. Correct Classification of group configurations of Meeting 1 and Meeting 2 with HMMs trained on group configurations of Meeting 2 (training set: Meeting 2, test set: Meeting 1).

States	*1*	*2*	*4*	*8*
Meeting 1	0.49	0.40	0.27	0.29
Meeting 2	0.77	0.35	0.26	0.21

Classification results of this approach are not very promising. In particular, the optimal state number for both training sets when classifying test and training seems to be one state. This indicates that the sequential structure of the observations, i.e. who speaks after whom, is not discriminating for the different group configurations. It is rather the distribution of the different observations, i.e. the number of interruptions, monologues of different participants, that seem to discriminate a group configuration. This is also due to the fact that we want to focus on short-term group configurations. In our meeting recordings, the duration of group configurations was between two and three minutes, which may not be sufficient for training and recognition of sequential speech patterns.

3.3 HMM based on Conversational Hypotheses

The second approach is to construct a Hidden Markov Model based on conversational hypotheses. These conversational hypotheses are translated to probability distributions of the different observations generated by the states of the HMM. These states correspond to the different group configurations. We can use the Viterbi algorithm [12] to calculate the most probable state sequence matching a given observation sequence, i.e. a given sequence of speech detection. This state sequence corresponds to the sequence of group configurations that have been recognized for the given observations. In the following, we will detail the conversional hypotheses as well as the construction of the Hidden Markov Model.

3.3.1 Conversational Hypotheses

This approach is based on basic conversational hypotheses. When two individuals are speaking at the same time, it is unlikely that they are in the same group. If two individuals do not speak, we can not decide their group membership. Finally, if one of the two individuals is speaking, it is likely that they are in the same group. We benefit from the regulation of speech activity, taking into account that verbal interaction are ordered within each group and disordered between different groups. As explained in Section 1, we assume that a group consists of at least two individuals.

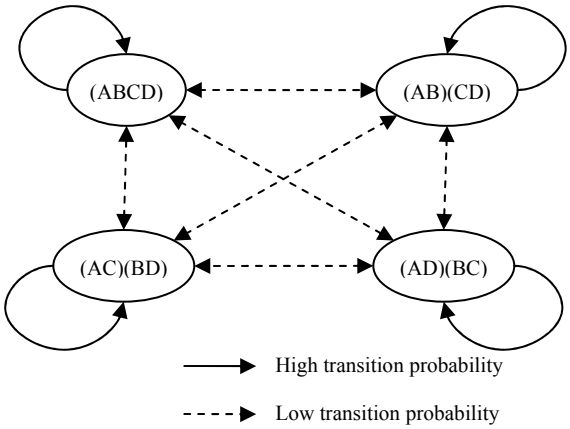

Figure 1. States of a HMM describing possible group configurations for a meeting of 4 individuals.

3.3.2 Construction of Hidden Markov Model

The construction of the Hidden Markov Model relies on the estimation of different probability distributions for the observations depending on the group configuration. Each possible group configuration is represented by a state of the HMM. And each state has a probability distribution for the observations associated. This probability distribution is based on the conversational hypotheses. We assume that the probability that two individuals of the same group are speaking at the same time is low, while this probability is high when two individuals are not in the same group. When all individuals form one big group, the probability that one single individual speaks is high, while the probability of several individuals speaking in parallel is low. The transition probabilities between the states are set to a very low level in comparison to the probabilities to stay in the same state of the HMM. This is necessary to stabilize the detection of state changes as the frequency of incoming observations (speech activity events) is very high in comparison to the dynamics of group changes (16ms compared to circa 30 seconds). We assume hence that group changes occur in reasonable delays. Figure 1 shows the states of a HMM for 4 individuals.

To detect the group configurations in real-time, we apply the Viterbi algorithm to the flow of arriving observations. Viterbi calculates the most probable state sequence that generated the observations. This state sequence corresponds to the sequence of detected group configurations. We calculate the state sequence for a window of the last 5000 arriving observation, which corresponds to an observation window of 80 seconds. The state calculated for the last observation represents the actual group configuration. The Viterbi algorithm is recalculated every 10 observations, which corresponds to a displacement of the window of 10 observations and an estimation of the group configuration every 160 milliseconds.

4. EVALUATION AND RESULTS

In this section, we describe the evaluation of the approaches. We recorded the interactions of 4 individuals during 3 experiments. The number and order of group configurations, i.e. who will speak with whom, was fixed in advance for the experiments. The

exact timestamps and durations of the group configurations were, however, not predefined and changed spontaneously. The individuals were free to move and to discuss any topic.

Figure 2. Picture of an example configuration of 2 independent groups of 2 individuals.

The speech of each individual was recorded using a lapel microphone. The speech activity detector [17] is executed on the audio channels of the different lapel microphones. We admit the use of lapel microphones in order to minimize correlation errors of speech activity of different individuals, i.e. speech of individual A is detected as speech of individual B. Figure 2 shows a picture of a configuration of 2 independent groups of 2 individuals during the experiments.

Table 4. Confusion matrix for the 3 experiments

Group Conf.	(ABCD)	(AB)(CD)	(AC)(BD)	(AD)(CB)
(ABCD)	**0.87535**	0.02737	0.06245	0.03460
(AB)(CD)	0.08535	**0.86707**	0.04549	0.00207
(AC)(BD)	0.21531	0.01057	**0.77411**	0.0
(AD)(BC)	0.03926	0.03629	0.07951	**0.84492**

4.1 Results

Table 4 shows the confusion matrix for the 3 experiments. This matrix indicates for each group configuration the correct and wrong detections. The lines of the matrix contain the detection results, while the columns contain the expected response.

We obtain a total recognition rate for the group configurations of 84.8 %. Figures 3, 4 and 5 give an overview of the detection of group configurations over time. The lines of the chart correspond to different group configurations. The continuous line indicates the correct group configuration expected as detection result.

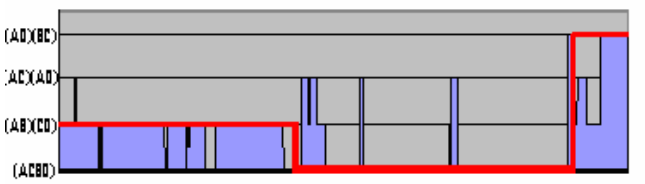

Figure 3. Group configuration detection during Experiment 1 (duration: 9 min. 22 sec.).

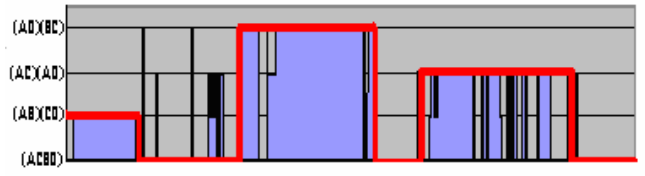

Figure 4. Group configuration detection during Experiment 2 (duration: 15 min. 16 sec.).

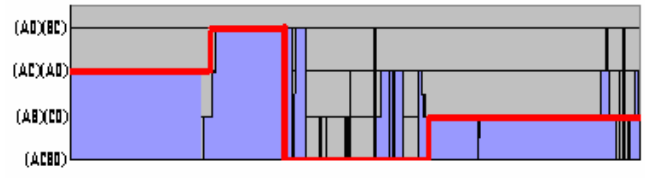

Figure 5. Group configuration detection during Experiment 3 (duration: 16 min. 19 sec.).

The results are encouraging and tend to validate the conversational hypotheses to distinguish interaction groups. The Viterbi algorithm executed on long observation sequences (like the observation window) is quite robust to wrong detections of the speech activity detector. However, a minimum number of correct speech activity detections is necessary, as the method relies on the information of who speaks at which moment. The use of lapel microphones made it possible to limit wrong detections as these microphones are attached to a particular person (and thus should only detect his/her speech).

5. CONCLUSION

We proposed a real-time detector for configurations of interaction groups. This detector is based on a HMM constructed upon conversational hypotheses. The input of the detector is a speech activity vector containing the information which individual is speaking or not. The synchronization of speech contributions within a group enables the detection of possible group configurations by a HMM built upon conversational hypotheses. The obtained results are encouraging, in particular as the group detection is exclusively based on speech activity, in presence of wrong speech activity detections.

The integration of further information into the model will be an important aspect of future research. Speech activity detection is not sufficient to disambiguate all situations, in particular to detect isolated individuals. Further information like head orientation and interpersonal distances seem to be good indicators [2]. Thus a multimodal approach needs to be envisaged.

Further audio recordings need to be done in order to validate and refine the conversational hypotheses. A big amount of representative conversational data will enable learning and adjustment of the probabilities of the HMM, which may improve its general performance. In addition, enough conversational data representing a specific context may permit to adapt the detector to a particular context to improve its performance in this context.

6. REFERENCES

[1] Basu S. *Conversational Scene Analysis.* Ph.D. Thesis. MIT Department of EECS. September, 2002.

[2] Beattie, G. *The regulation of speaker turns in face-to-face conversation;some implications for conversation in sound-only communication channels*, Semiotica, 34, 55-70, 1981

[3] Burger, S., MacLaren, V., and Yu, H., *The ISL Meeting Corpus: The Impact of Meeting Type on Speech Style,* Proc. of ICSLP 2002, Denver, CO, USA, 2002.

[4] Choudhury, T. and Pentland, A., *Characterizing Social Interactions Using the Sociometer*, Proceedings of NAACOS 2004, June 2004

[5] Clark H. *Using Language*. Cambridge University Press, 1996.

[6] Clark H., Schaefer E., F. *Contributing to Discourse*, Cognitive Science 13, 259-294, 1989.

[7] Duncan, S., *Some signals and rules for taking speaking turns in conversation*, Journal of Personality and Social Psychology, 23(2), 283-293, 1972.

[8] Goffman, E. *Frame analysis*. Harper Row, New York, 1974.

[9] Goffman, E. *Footing.* Forms of Talk, 124-159, University of Pennsylvania Press, Philadelphia, 1981.

[10] Joseph, I., *Attention distribuée, attention focalisée. Les protocoles de la coopération au PCC de la ligne A du RER*. Sociologie du Travail, 4/94, 563-587, 1994.

[11] McCowan, I., Gatica-Perez, D., Bengio, S., Lathoud, G., Barnard, M., and Zhang, D., *Automatic Analysis of Multimodal Group Actions in Meetings,* IEEE Transactions on Pattern Analysis and Machine Intelligence, 2004.

[12] Rabiner, L. *A tutorial on Hidden Markov Models and selected applications in speech recognition*, Proc. IEEE 77(2):257-286, 1989.

[13] Sacks, H., Schegloff, E., and Jefferson, G. *A simplest systematics for the organization of turn-taking for conversation*. Language, 50, 696-735, 1974.

[14] Huang, X.D., Ariki, Y., and Jack, M.A. *Hidden Markov Models for Speech Recognition.* Edinburgh University Press, 1990.

[15] Starner, T.E. *Visual Recognition of American Sign Language Using Hidden Markov Model.* PhD Thesis, MIT, Media Laboratory, Perceptual Computing Section, 1995.

[16] Martin, J., and Durand, J.-B. *Automatic Handwriting Gestures Recognition Using Hidden Markov Models.* FG 2000: 403-409.

[17] Vaufreydaz, D., *IST-2000-28323 FAME: Facilitating Agent for Multi-Cultural Exchange (WP4)*, European Commission project IST-2000-28323 October 2001.