Unsupervised Segmentation of Meeting Configurations and Activities using Speech Activity Detection

Oliver Brdiczka, Dominique Vaufreydaz, Jérôme Maisonnasse, Patrick Reignier

Unsupervised Segmentation of Meeting Configurations and Activities using Speech Activity Detection

Oliver Brdiczka, Dominique Vaufreydaz, Jérôme Maisonnasse, Patrick Reignier

INRIA Rhône-Alpes
655 Av. de l'Europe
38330 Montbonnot, France
{brdiczka, vaufreydaz, maisonnasse, reignier}@inrialpes.fr

This paper addresses the problem of segmenting small group meetings in order to detect different group configurations and activities in an intelligent environment. Our approach takes speech activity detection of individuals attending a meeting as input. The goal is to separate distinct distributions of speech activity observation corresponding to distinct group configurations and activities. We propose an unsupervised method based on the calculation of the Jeffrey divergence between histograms of speech activity observations. These histograms are generated from adjacent windows of variable size slid from the beginning to the end of a meeting recording. The peaks of the resulting Jeffrey divergence curves are detected using successive robust mean estimation. After a merging and filtering process, the retained peaks are used to select the best model, i.e. the best speech activity distribution allocation for a given meeting recording. These distinct distributions can be interpreted as distinct segments of group configuration and activity. To evaluate, we recorded 6 small group meetings. We measured the correspondence between detected segments and labeled group configurations and activities. The obtained results are promising, in particular as our method is completely unsupervised.

Introduction

Ubiquitous computing [14] integrates computation into all-day environments. People are enabled to move around and interact with computers more and more naturally. One of the goals of ubiquitous computing is to enable devices to sense changes in the environment and to automatically adapt and act based on these changes. A main focus is laid on sensing and responding to human activity. Human actors need to be identified in order to perceive correctly their activity. In order that ubiquitous computer devices act and interact correctly with users, addressing the right user at the correct moment and perceiving his correct activity is essential. Thus we need to detect potential users and their connection while they are doing an activity.

The focus of this work is analyzing human (inter)action in meeting environments. In these environments, users are collaborating in order to achieve a common goal. 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 and activity need to be tracked in order to supply reactions or interactions at the most appropriate moment. Changes in group configuration need to be detected to identify main actors, while changes in activity within a group need to be detected to identify activities.

This paper proposes an unsupervised method for detecting changes in group configuration and group activity based on measuring the Jeffrey divergence between adjacent histograms.

These histograms are calculated for a window sliding from the beginning to the end of the meeting and contain the frequency of (human) activity events. The peaks of the Jeffrey divergence curve are used to segment distinct distributions of activity events and to find the best model of activity event distributions for the given meeting. The method has been tested on speech activity detection events as sensor information for interacting individuals. We focus thus on verbal interaction. The evaluation has been done with speech activity recordings of 6 meetings.

Previous and Related Work

Many approaches for the recognition of human activities in meetings have been proposed in recent years. Most work uses supervised learning methods [2], [5], [7], [11]. Some projects focus on supplying appropriate services to the user [11], while others focus on the correct classification of meeting activities [5] or individual availability [7]. Less work has been conducted on unsupervised learning of meeting activities [15]. 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 [5] 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. The automatic detection of conversations using mutual information [1], in order to determine who speaks and when, needs an important duration of each conversation. To our knowledge, little work has been done on the analysis of changing small group configuration *and* activity. In [2] a real-time detector for changing small group configurations has been proposed. This detector is based on speech activity detection and either trained with recorded meetings or defined by hand based on conversational hypotheses. In [2], the authors showed that different meeting activities, and especially different group configurations, have particular distributions of speech activity. Detecting group configuration or activity [2], [5], [7] requires, however, a predefined set of activities or group configurations. New activities or group configurations with a different number of individuals cannot be detected and distinguished with these approaches. Our approach focuses on an unsupervised method segmenting small group meetings into consecutive group configurations and activities. These configurations and activities are distinguished by their distributions, but not labelled or compared. The method can thus be seen as a first step within a classification process identifying (unseen) group configurations and activities in meetings.

Approach

Speech Activity Detector: A Multi-Agent System

Our approach is based on speech activity detection (SAD) of individuals attending a meeting. We are recording the speech of each individual using lapel microphones. We admit the use of lapel microphones in order to minimize detection errors. An automatic speech detector parses the audio channels of the different lapel microphones and detects which individual stops and starts speaking.

Our speech activity detector is composed of several sub-systems: an energy detector, a basic classifier and a neural net trained to recognize voiced segments like vowels for example. At each time, i.e. for each frame, each sub-system gives an answer indicating whether the input

signal is speech or not. A hand-crafted rule based automaton then determines the final result: speech activity or not. The complete system is shown in Figure 1.

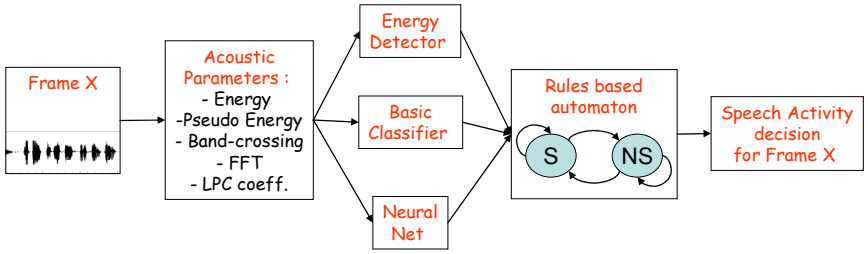


Figure 1: Diagram of our SAD system.

This speech activity detector is designed to be light-weight; it can be run efficiently on more than 50 channels at the same time. The output 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 for further treatment. The automatic speech activity detector generates an output observation every 16 milliseconds.

Energy detector

The energy detector uses pseudo energy computation in order to detect variations of the input signal energy. Its answer can be START_SPEAKING, STILL_SPEAK, STOP_SPEAK or SILENCE. The detection is based on couples of time delays and thresholds: *TimeOn*/*EnergyOn* and *TimeOff*/*EnergyOff*. The initial values for time delays and thresholds were the ones used during the NESPOLE! Project [6]. The energy detector is then able to adapt dynamically *EnergyOn* and *EnergyOff,* given the final SAD answer to retrain itself. *TimeOn* et *TimeOff* were set to 100 ms and 800 ms respectively.

Basic Classifier

This classifier is dedicated to recognize and to tag three specific sound classes: fricatives, low frequency sounds like computer or air conditioning fans, and all other sounds. The classifier computes the energy for 5 identical sub-frequency bands on the spectrum from 1 to 8000 hertz (higher frequencies are not considered). Given the 5 energy values, the module classifies the audio signal.

Neural Net

The neural net is a multi-layer perceptron with 2 hidden layers. It uses advanced coefficients computed on the input frames as input: *band-crossing* [12], *Energy and* 16 *predictor coefficients* extracted from a speech analysis method called Linear Predictive Coding (LPC) [10]. This module is the only sub-system that needs to be trained. The training was made on 1 hour of French speech extracted from the BREF corpus [3]. The phonetic labels used during the training phase are not the original BREF ones but were computed with RAPHAEL [13], a French recognizer.

Precision/Recall

In the following table, we summarize our SAD accuracy using the annotated evaluation data from the CHIL project [11] in the same experimental conditions.

Table 1. SAD results calculated within the CHIL evaluations.

	Recall [%]	**Precision** [%]	**Fallout** [%]	**Error** [%]
Lapel micropone	96.45	89.10	37.72	11.46

Speech Activity Distributions

In [2], the authors stated that the distribution of the different speech activity observations is discriminating for group configurations in small group meetings. Thus we assume that in small group meetings distinct group configurations and activities have distinct distributions of speech activity observations. The objective of our approach is hence to separate these distinct distributions, in order to identify distinct small meeting configurations and activities.

The observations of the speech activity detector are an unordered 1-dimensional discrete code indicating who is currently speaking (e.g. for four lapel microphones, the code is between 0 (no speech) and 15 (everybody is speaking)). As we do not want to admit any a priori distribution, we use histograms to represent speech activity distributions. A histogram is calculated for an observation window (i.e. the observations between two distinct time points in the meeting recording) and contains the frequency of each observation code within this window.

To separate different speech activity distributions, we calculate the Jeffrey divergence [8] between the histograms of two adjacent observation windows. The Jeffrey divergence is a numerically stable and symmetric form of the Kullback-Leibler divergence between histograms. We slide two adjacent observation windows from the beginning to the end of the recorded meetings, while constantly calculating the Jeffrey divergence between these windows. The result is a divergence curve of adjacent histograms (Figure 2).

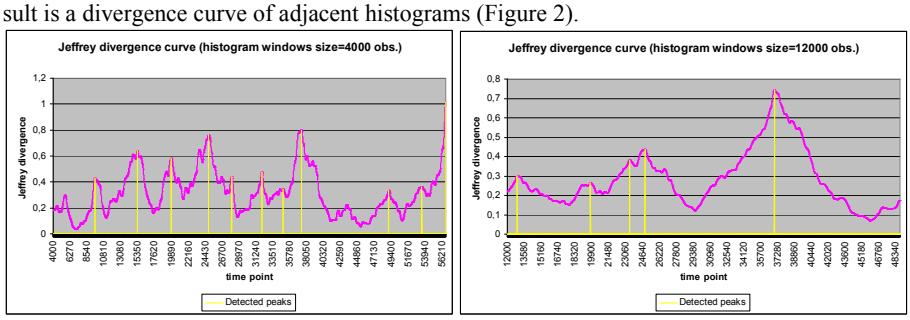


Figure 2. Meeting 5: Jeffrey divergence between histograms of sliding adjacent windows of 4000, and 12000 observations (64sec and 3min 12sec)

The peaks of the curves indicate high divergence values, i.e. a big difference between the adjacent histograms at that time point. The size of the adjacent windows determines the exactitude of the divergence measurement. The larger the window size, the less peaks has the curve. However, peaks of larger window sizes are less precise than those of smaller window sizes. Thus we parse the meeting recordings with different window sizes (sizes of 4000, 6000, 8000, 10000, 12000, 14000 and 16000 observations, which corresponds to a duration between 64sec and 4min 16sec for each window). The peaks of the Jeffrey divergence curve can then be used to detect changes in the speech activity distribution of the small meeting recording.

Peak Detection

To detect the peaks of the Jeffrey divergence curve, we use successive robust mean estimation. Robust mean estimation has been used in [9] to locate the center position of a dominant face in skin color filtered images. Mean and standard deviation are calculated repeatedly in order to isolate a dominant peak. To detect all peaks of the Jeffrey divergence curve, we apply the robust mean estimation process successively to the Jeffrey divergence values.

Merging and Filtering Peaks from different Window Sizes

Peak detection using successive robust mean estimation is conducted for Jeffrey curves with histogram window sizes of 4000, 6000, 8000, 10000, 12000, 14000 and 16000 observations. A global peak list is maintained containing the peaks of different window sizes. Peaks in this list are merged and filtered with respect to their window size and peak height.

The small number of peaks resulting from merging and filtering is used to search for the best allocation of speech activity distributions, i.e. to search for the best model for a given meeting.

Model Selection

To search for the best model for a given meeting recording, we examine all possible peak combinations, i.e. each peak of the final peak list is both included and excluded to the (final) model. For each such peak combination, we calculate the average Jeffrey divergence of the histograms between the peaks. As we want to separate most distinct speech activity distributions, we accept the peak combination that maximizes the average divergence between the peak histograms as the best model for the given meeting.

EVALUATION AND RESULTS

The result of our approach is the peak combination separating best the speech activity distributions of a given meeting recording. As we admit that distinct distributions of speech activity are discriminating for group configurations and activities in small group meetings [2], we interpret the intervals between the peaks as segments of distinct group configuration and activity. To evaluate our approach, we recorded 6 small group meetings. The group configurations and activities of these meetings have been labeled. For the evaluation of the detected segments, we use the *asp*, *aap* and *Q* measures proposed in [15].

Experiments

To evaluate our approach, we recorded 6 small group meetings (between 4 and 5 individuals). The number and order of group configurations, i.e. who will speak with whom, and of group activities, e.g. presentation/questions/discussion etc., were fixed in advance for the experiments. The timestamps and durations of the group configurations and activities were, however, not predefined and changed spontaneously. The individuals were free to move and to discuss any topic.

Evaluation measures

To evaluate, we dispose of the timestamps and durations of the (correct) group configurations and activities. However, classical evaluation measures like confusion matrices can not be used here because the unsupervised segmentation process does not assign any labels to the found segments.

$$asp = \frac{1}{N}\sum_{i=1}^{N_s} p_{i\bullet} \times n_{i\bullet} \quad , \quad aap = \frac{1}{N}\sum_{j=1}^{N_a} p_{\bullet j} \times n_{\bullet j} \quad , \quad Q = \sqrt{asp \times aap}$$

with

n_{ij} = *total number of observations in segment i by activity j*

$n_{i\bullet}$ = *total number of observations in segment i*

$n_{\bullet j}$ = *total number of observations*

N_a = *total number of activities*

N_s = *total number of segments*

N = *total number of observations*

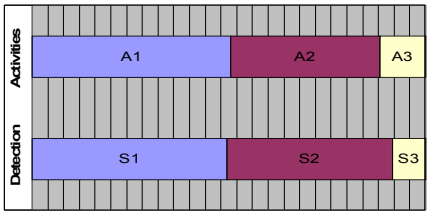
$$p_{i\bullet} = \sum_{j=1}^{N_a} \frac{n_{ij}^2}{n_{i\bullet}^2} \qquad p_{\bullet j} = \sum_{i=1}^{N_s} \frac{n_{ij}^2}{n_{\bullet j}^2}$$

Figure 3. Average segment purity (*asp*), average activity purity (*aap*) and the overall criterion *Q*

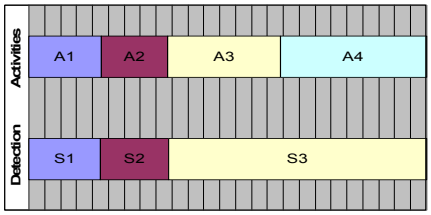
Instead, we use three measures proposed in [15] to evaluate the detection results: average segment purity (*asp*), average activity purity (*aap*) and the overall criterion *Q* (Figure 3). The *asp* is a measure of how well a segment is limited to only one activity, while the *aap* is a measure of how well one activity is limited to only one segment. The *Q* criterion is an overall evaluation criterion combining *asp* and *aap*, where larger *Q* indicates better overall performance.

Results

Figure 4 shows the labeled group configurations/activities for each small group meeting as well as the segments detected by our approach. Table 2 indicates the *asp*, *aap* and *Q* values for each meeting as well as the average of these values for all meetings. Unlike meeting recordings 1, 4, 5 and 6, recordings 2 and 3 contain numerous wrong speech activity detections caused by correlation errors and microphone malfunctions. However, our approach worked well for meeting recording 2, while the segmentation of meeting recording 3 is mediocre. The overall results of our approach are very good; the average *Q* value is 0.82. By excluding meeting 3, we even obtain a *Q* value of 0.88.



Meeting 1 (Q=0.93, duration=9min 14sec)



Meeting 2 (Q=0.81, duration=10min 14sec)

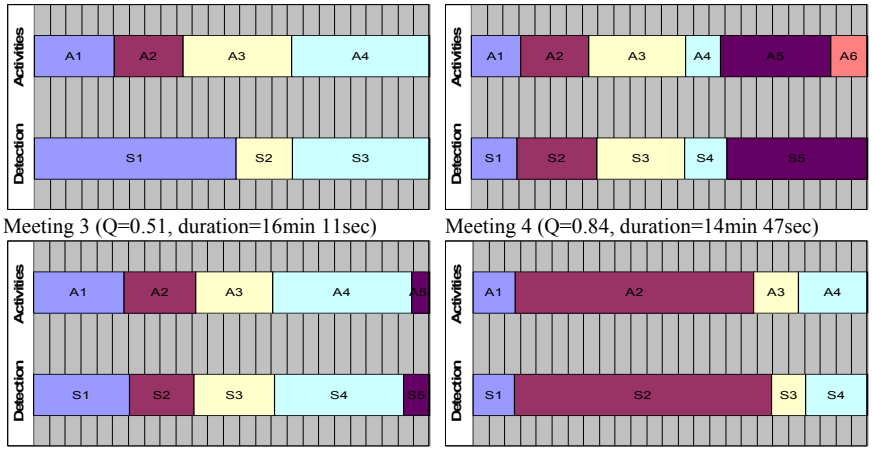


Figure 4. Group Configurations/Activities and their detection for meetings 1-6.

Table 2. *asp*, *aap* and *Q* values for the recorded meetings.

	asp	**aap**	**Q**
Meeting 1	0.93	0.92	0.93
Meeting 2	0.67	0.99	0.81
Meeting 3	0.42	0.62	0.51
Meeting 4	0.78	0.91	0.84
Meeting 5	0.92	0.91	0.92
Meeting 6	0.88	0.91	0.90
Average	0.77	0.88	0.82

CONCLUSION

We proposed an unsupervised method for segmenting small group meeting configurations and activities. This method is based on the calculation of the Jeffrey divergence between histograms of observations of speech activity. The peaks of the Jeffrey divergence curve are used to separate distinct distributions of speech activity observations. These distinct distributions can be interpreted as distinct segments of group configuration and activity. We measured the correspondence between the detected segments and labeled group configurations and activities. The obtained results are promising, in particular as our method is completely unsupervised.

Further meeting recordings need to be done in order to apply and evaluate our method on more and subtler meeting activities. These meeting activities will include activity changes within a group configuration.

Our method can help obtaining a first segmentation of a meeting. The detected segments can then be used as input for further classification tasks like meeting comparison, meeting activity recognition etc.

Future work will concern the test of our method on further meeting information. Speech activity detection is not sufficient to disambiguate all situations. Further information like head orientation, pointing gestures or interpersonal distances seem to be good indicators. Thus a multimodal approach needs to be envisaged. The method can easily be extended to such an approach as we only need to upgrade the observation codes used for the generation of the histograms.

References

1. Basu S., *Conversational Scene Analysis,* Ph.D. Thesis. MIT Department of EECS. September, 2002.
2. Brdiczka, O., Maisonnasse, J., and Reignier, P., *Automatic Detection of Interaction Groups,* Proc. Int'l Conf. Multimodal Interfaces, 2005 (to appear).
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. Lamel L., Gauvain J.L., Eskenazi M., *BREF, a large vocabulary spoken corpus for French,* Eurospeech'91, Gênes (Italie), 1991.
5. McCowan, I., Gatica-Perez, D., Bengio, S., Lathoud, G., Barnard, M., and Zhang, D., *Automatic Analysis of Multimodal Group Actions in Meetings,* IEEE Trans. on Pattern Analysis and Machine Intelligence, vol. 27, no. 3, pp. 305-317, March 2005.
6. Metze, F., Mc Donough, J., Soltau, H., Waibel, A., Lavie, A., Burger, S., Langley, C., Levin, L., Schultz, T., Pianesi, F., Cattoni, R., Lazzari, G., Mana, N., Pianta, E., Besacier, L., Blanchon, H., Vaufreydaz, D., Taddei, L., *The Nespole! Speech-to-Speech Translation System*, Human Language Technologies 2002, San Diego, California (USA), March 2002.
7. Muehlenbrock, M., Brdiczka, O., Snowdon, D., and Meunier, J.-L., *Learning to Detect User Activity and Availability from a Variety of Sensor Data,* Proc. IEEE Int'l Conference on Pervasive Computing and Communications, March 2004.
8. Puzicha, J., Hofmann, Th., and Buhmann, J., *Non-parametric Similarity Measures for Unsupervised Texture Segmentation and Image Retrieval*. Proc. Int'l Conf. Computer Vision and Pattern Recognition, 1997.
9. Qian, R. J., Sezan, M. I., and Mathews, K. E., *Face Tracking Using Robust Statistical Estimation*, Proc. Workshop on Perceptual User Interfaces, San Francisco, 1998.
10. Rabiner L., Juang B.H., *Fundamentals of Speech Recognition*, Prentice Hall PTR, ISBN 0-130-15157-2, 1993.
11. Stiefelhagen, R., Steusloff, H., and Waibel, A., *CHIL - Computers in the Human Interaction Loop*, Proc. Int'l Workshop on Image Analysis for Multimedia Interactive Services, 2004.
12. Taboada J., Feijoo S., Balsa R., Hernandez C., *Explicit estimation of speech boundaries*, IEEE Proc. Sci. Meas. Technol., vol. 141, pp. 153-159, 1994.
13. Vaufreydaz, D., *Modélisation statistique du langage à partir d'Internet pour la reconnaissance automatique de la parole continue*, Ph.D. thesis in Computer Sciences, University Joseph Fourier, Grenoble (France), 226 pages, January 2002.
14. Weiser, M., *Ubiquitous Computing: Definition 1*, http://www.ubiq.com/hypertext/weiser/UbiHome.html, March 1996.
15. Zhang, D., Gatica-Perez, D., Bengio, S., McCowan, I., and Lathoud, G., *Multimodal Group Action Clustering in Meetings***,** Proc. Int'l Workshop on Video Surveillance & Sensor Networks, 2004.