Voice activity detection from gaze in video mediated communication

Michal Hradis*
Brno University of Technology

Shahram Eivazi†
Roman Bednarik‡
University of Eastern Finland

Abstract

This paper discusses estimation of active speaker in multi-party video-mediated communication from gaze data of one of the participants. In the explored settings, we predict voice activity of participants in one room based on gaze recordings of a single participant in another room. The two rooms were connected by high definition, low delay audio and video links and the participants engaged in different activities ranging from casual discussion to simple problem-solving games. We treat the task as a classification problem. We evaluate several types of features and parameter settings in the context of Support Vector Machine classification framework. The results show that using the proposed approach vocal activity of a speaker can be correctly predicted in 89 % of the time for which the gaze data are available.

CR Categories: I.2.m [Artificial Intelligence]: Miscellaneous;

Keywords: gaze tracking, voice activity detection, machine learning, Support Vector Machines, video-mediated communication

1 Introduction

Eye gaze[1] is central for grounding during communication in that gaze signals are important for collecting and providing information for mutual understanding [Clark and Brennan 1991]. While it is well established that eye-movements are a good proxy to the allocation of attention [Rayner 1998], during conversation, eye-movements also carry the information about how well the interlocutors understand each other [Richardson et al. 2007]. Without eye contact, it is hard to engage in an efficient conversation [Argyle and Cook 1976].

In systems supporting multi-party video-mediated (MPVM) communication, a principal problem is presenting information from a remote location on a limited visualisation device. This challenge has to be solved by composing the information in the available screen-space and time in an appropriate way. The systems have to be able to present the remote information compactly on the screen [Jansen et al. 2011] which would be ideally achieved by performing automatic directorial decisions in real-time [Falelakis et al. 2011; Ursu et al. 2011] based on inferred information about current activity of the participants and current interaction state.

Clearly, such systems are not available at the moment for several reasons. One reason is that for the directorial decisions to intelligently and effectively aid communication, diverse knowledge from several disciplines has to be combined. The involved areas of research include, for example, sensor-based systems, computer vision, machine learning, social psychology and cinematics. It turns out that one of the required aspects is a deep understanding of attention during collaboration and communication.

In multi-party interaction, certain information can be discarded without negative influence on understandability and naturalness of the conversation, while omitting other information can make the interaction incomprehensible or frustrating [Ursu et al. 2011]. In this paper we aim to broaden our understanding of attention in multi-party video-mediated communication by exploring the link between gaze and speech.

We investigate the hypothesis that voice activity of participants in multi-party mediated communication can be estimated from gaze of a listener. The explored task is to estimate the voice activity – who is speaking and when[2]– of several participants simultaneously located in a single room. We carry out such analysis only based on gaze information recorded for a single remote participant.

We designed a study in which a group of participants had a conversation with another participant, remotely connected by a high-definition low latency audio and video link. Such setup is common for example in business meetings, remote assistance, or online lecturing. The participants had known each other prior to the recordings and the recorded activities range from natural discussion about casual events to simple problem-solving games. Although the presented task is interesting by itself and it could find applications in real-time communication, we hope that this study will presents valuable insight into attention of participants in MPVM interaction.

The approach we chose is based on learning discriminative Support Vector Machine (SVM) classifiers which estimate voice activity of a participant based on a feature vector extracted from a fixed time-window of gaze data. We present several types of features, their results on the dataset and analysis of the recorded gaze data.

1.1 Gaze in multi-party communication

Understanding speaker activities during MPVM communication have important implications on designing any system that is able to proactively coordinate or structure communications. There are various non verbal means for detecting speaker activities implicitly. For example, analyzing the speaker head position, gaze, facial expression, gestures. Speakers use effectively gestures, facial expressions, and body posture signals to coordinate their communicative activity in conversations [Jokinen 2009; Jokinen et al. 2010]. [Rienks et al. 2010] shows how people recognize the speaker from listeners using patterns of head orientation. However, their result for speakers identification was only 43.27 % on average, which suggests that head orientation information alone is not sufficient for predicting speakers in multi-party settings.

While a speaking interlocutor is likely to attract attention of the listeners in some way, little is known about the details of this process. [Griffin and Bock 2000] explored the time course between fixation and spoken word. Their observations show that speakers fixation

*e-mail:ihradis@fit.vutbr.cz
†e-mail:seivazi@cs.joensuu.fi
‡roman.bednarik@uef.fi

[2] Voice activity is understood as any verbal and nonverbal vocal activity.

point on an object less than one second before the speaker would talk about the object. Recently [Jokinen et al. 2009; Jokinen et al. 2010] collected gaze data from speakers in a natural dialogues setting. They report that gaze data plays an important role as a signal to define who has potential to be the next speaker. However, according to their findings, gaze and speech are more often parallel and not complimentary sources of communication signals.

2 Method

We selected a classification approach to the speaker estimation task which is based on SVM classifiers. Feature vectors are extracted for a participant from a constant-length time-windows on a gaze recording of one participant. The features are computed from fixations and saccades in a way which takes into account head position of the participant for which the feature vector is created as well as the head positions of other participants. The task of the classifier is to decide whether the person for which the feature vector was extracted is speaking or not. The decision is made independently for each participant and thus any number of them can be predicted to speak simultanously.

The feature vectors are extracted in a way which is not explicitly dependent on the number of participants and thus the created classifiers can be used without any change in sessions with different number of participants.

2.1 Feature Extraction

For the feature extraction, fixations and saccades were detected in the raw gaze data by a velocity-based algorithm [Salvucci and Goldberg 2000] for fixation identification with a threshold 100 deg/sec, followed by a local dispersity based identification. A minimum temporal threshold for fixation duration was 100 ms and the maximum distance between two gaze points belonging to the same fixation was set to 40 pixels. This hybrid algorithm and settings performed best on the recorded dataset when manually compared with other traditional approaches.

Altogether we used ten different features extracted from the gaze data to train the classifiers. The features can be divided into three groups: The first group includes features giving information about fixations that are inside a ground truth bounding box of the head of the participant for whom the feature vector is computed. The second group includes features providing information about fixations on head positions of all the other participants in the room. Features in the last group describe fixations outside the head positions of any of the participants and can be thought of as fixations on objects in the room or other fixations which do not correspond to attention of the viewer on the participants.

The features in each of the three groups, were number of fixations, average of fixation durations, and average of saccade lengths (an euclidean distance between two consecutive fixations).

In addition to these nine features, the tenth feature was computed - a number of participants that have been visually attended by the viewer in the examined time-window.

As mentioned before, features were extracted for fixed-length time-window. The time-windows utilized information only from the past - the ground truth labels for classification were set according whether the corresponding person is speaking at the moment the time-window ends. This type of feature extraction, which does not use future information, fits well for time-critical real-time systems as it provides zero latency (except for the time needed for computing the classification function).

In general, it is an open question what length of the time-windows should be used for a particular domain and problem. The optimal length would probably be different for different scenarios. Longer windows provide more contextual information which we expect is useful for the considered classification task. On the other hand, long windows make localization of short utterances less precise. The opposite is true for short windows which provide very little information while allowing very precise utterance localization. In an attempt to utilize the context information, as well as the localization information, we extracted features from windows of different lengths aligned such that they end the same time. Feature vectors for classification were then created simply by concatenating the features from the individual time-windows into a single feature vector.

2.2 Classification

For classification, each feature was normalized by linear transformation into an interval [0,1]. An SVM with linear and Gaussian kernel was employed to create the classifiers. We used LIBSVM implementation of a solver for the standard soft-margin SVM formulation [Cortes and Vapnik 1995] which has a single regularization parameter C. We used SVM as it is a standard stat-of-the-art of-the-shelf classifiers and other discriminative classifiers could be probably used with similar results.

Optimal value of the SVM regularization parameter C was estimated by 2D grid search together with the parameter γ of the Gaussian Kernel

$$K(x, x') = exp\left(-\gamma\|x - x'\|_2^2\right), \qquad (1)$$

where x and x' are feature vectors.

The objective function in the hyper-parameter optimization was Equal Error Rate (EER). The EER for certain hyper-parameter setting was estimated by a cross-validation where the folds were data from the individual gaze-recording sessions. This way the classifiers were prevented to utilize any knowledge specific to the testing session during the learning phase. This is consistent with the concluded experiments where performance across different sessions with different participants was evaluated. In the experiments, the performance measures were estimated by a second level of cross-validation with the folds equal to the recording sessions as in the case of hyper-parameter optimization.

3 Dataset

The corpus was recorded in two separate rooms. Room A contained a living-room table, sofa and chairs. Only one office chair was placed in Room B in front of a desk with a display and speakers. The two rooms were connected by low-delay, high-quality audio and video links. Person in Room B was shown an wide-angle frontal view of Room A (as shown in Figure 1). The camera covered all participants in that room. The video was shown on a 24 inch screen at about 0.8 m from the viewer. The participants in Room A were shown a close-up view of the person in Room B (shown in Figure 2) also on a 24 inch screen located approximately 2.5 m in front of them. The delay of video was approximately 100ms and there was no perceivable audio delay.

During the sessions, audio and video data were recorded in parallel with the gaze data of the person in Room B. The cameras were Sony HDR-FX1E in Room A and Canon HV30 in Room B. Both cameras captured the video in 1080p@25 fps and we recorded directly the internally compressed mpeg2 stream. An array of four omnidirectional microphones (AKG C562CM) placed on the table in front of the participants was used to capture sound in Room A.

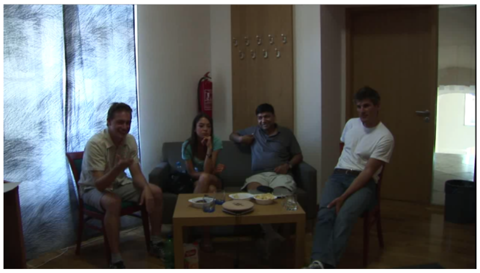

Figure 1: *The recorded view of Room A, showing the same view as was transmitted to the remote participant.*

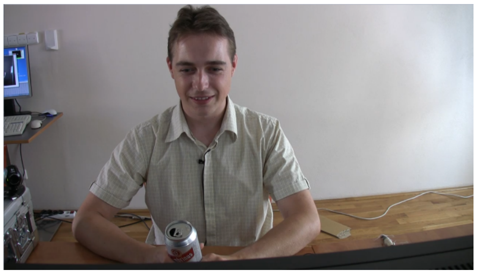

Figure 2: *The recorded view of Room B. It is the same view as was transmitted to the remote participants in Room A.*

The voice of the person in Room B was recorded using a clip-on microphone.

Eye movement data were recorded using Tobii X120 eye tracker (120Hz), at a viewing distance 60 cm. The participants within a group would take turns in front of the eye-tracker in Room B, often until everybody was at least once eye-tracked.

The volunteering participants were mostly master or doctoral students at one of the Brno universities and their friends from the same age group. The language of these sessions was Czech. Additionally, three researchers from other countries joined the recordings for one extra session which was in conducted then in English.

The complete recordings include 12 hours and 30 minutes of recorded video-material from each room and 28 gaze recordings with average length of 24 minutes (total 673 minutes).[3]

In this paper, we employ only a subset of the whole database. The details for sessions that provide the input dataset used in this study are shown in Table 1. The sessions to be included in the present analysis were chosen randomly and their number was constrained only by the speed of annotation process.

Duration	Viewer gender	males	females	fixations
30	male	4	1	1881
40	male	3	0	2709
47	male	3	0	3589
22	male	2	2	991
28	male	2	2	3129

Table 1: *Details of the recording sessions: the duration of the recording, gender of the viewer for which gaze was recorded, number of male and female participants in Room A, and the number of recorded fixations.*

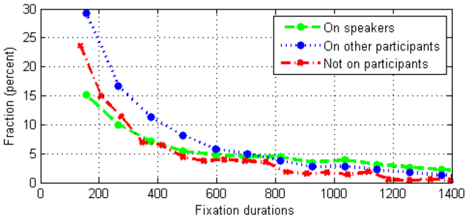

Figure 3: *Distribution of fixations when someone in Room A is talking. The figure shows distributions of fixations on active speakers, on not speaking participant and distribution of fixations which are not on the participants.*

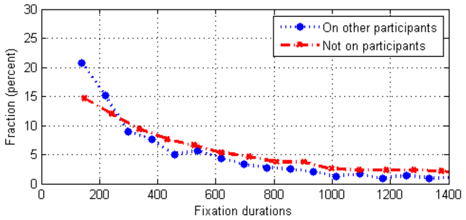

Figure 4: *Distribution of fixations when no-one in Room A is talking. The figure shows distributions of fixations on participants and of fixations that are not on the participants.*

Head positions were hand-annotated in each frame using ViperGT annotation tool. Voice activity was hand-annotated in ELAN, where the annotator could see the recorded video and the respective waveform, and hear the sound recorded in Room A. The annotations were created by two coders and subsequently checked by a experienced researcher. Any part of the recordings was annotated only by a single person.

A descriptive analysis of the fixations was divided into two parts. In the first case we analysed data when at least one participant was talking. The second case consisted of data when none of the participants was talking. The distributions of fixation durations for the two subsets are presented in Figures 3 and 4. The figures show that there are less short fixations on the speakers and that, on the other hand, fixations on not speaking participants tend to be shorter. Table 2 shows information about how often and for how long the gaze of the recorded participants stayed on active speakers, on the silent participants, and how long it stayed on other parts of the video image.

4 Experiments and Results

We evaluated the ability to automatically establish whether a participant is speaking from the extracted gaze features (see Section 2.1) in a way which is consistent with evaluation in other classification tasks. Each time point is classified independently. The reported error measure is Equal Error Rate (EER) which was estimated in cross-validation (see Section 2.2) on the available dataset. Because of computational reasons, the dataset was sub-sampled to a balanced set of 8000 samples. Only the samples with corresponding valid gaze data were considered.

As a baseline comparison, we considered a system predicting speakers directly according to the distance of current fixation to the participants. This baseline provided 28 % ERR.

When only a single time-window is used for feature extraction, its length significantly influences results. The results in Table 3 shows

[3] The dataset is available for download at http://medusa.fit.vutbr.cz/TA2/TA2/.

	Someone talking		No-one talking	
	Number of fixations (%)	Fixation durations (%)	Number of fixations (%)	Fixation durations (%)
Looking at speakers	39	60	-	-
Looking at other participants	38	26	36	31
Not looking at participants	21	12	63	68

Table 2: *Distribution of fixations and their durations between speakers, other participant and the rest of the room for situation when someone is speaking and for situation when no-one is speaking.*

Win. length [s]	0.2	0.4	0.6	0.8	1	1.2	1.4
Gaussian	18	22	25	24	24	24	24
Linear	23	24	26	25	25	28	27

Table 3: *Results for different window lengths and different kernels. EER is shown in percents.*

that localization provided by shorter windows is more important than context, and that better results are achieved by the short windows.

Further, we evaluated a system combining time-windows of lengths 200 ms, 400 ms, 600 ms and 800 ms (see Section 2.1). This resulted in an improved EER of 11 % for Gaussian kernel and 13 % for linear kernel.

5 Discussion and Conclusions

The importance of multi-party video-mediated communication will grow in the future as is indicated by the current trend of increasing installations of video communication services (e.g. Skype TV) in living rooms. Intelligent data analysis to predict the importance and activity during MPVM communication will become an important part of the systems allowing more natural interaction.

Our research is the first to show that eye movements alone are a good predictor of speaker activity in natural conversation. Our approach, using statistical machine learning methods, achieved 11 % EER for the task of estimating speaker activity from gaze of a single remote participant by combining features extracted from time-windows of different lengths aligned on the same position in time. The approach does not use future information, and is thus suitable for real-time applications.

While the proposed approach provides good results, it could be extended in several directions which should lead into further improvements in performance. The combination of features from different time-windows which was used is rather basic and methods that are able to estimate importance of features (such as multi-kernel learning) should give better results. Further, the presented approach does not consider interaction dynamics which can be modeled for example by Hidden Markov Models. It would also be interesting to combine the gaze data with information extracted from other modalities and aim at turn-taking prediction [Jokinen et al. 2010] instead of voice activity detection.

Acknowledgements

This work has been supported by the EU FP7 project TA2: Together Anywhere, Together Anytime ICT-2007-214793, grant no 214793; and by the research project Security-Oriented Research in Informational Technology CEZMŁMT, MSM0021630528. The work of R.B. has been supported by the Academy of Finland project # 137773.

References

Argyle, M., and Cook, M. 1976. *Gaze and mutual gaze.* Cambridge U Press, New York.

Clark, H. H., and Brennan, S. E. 1991. Grounding in Communication. In *Perspectives on socially shared cognition*, L. B. Resnick, J. M. Levine, and S. D. Teasley, Eds. American Psychological Association, 127–149.

Cortes, C., and Vapnik, V. 1995. Support-Vector Networks. *Machine Learning 20*, 3, 273–297.

Falelakis, M., Kaiser, R., Weiss, W., and Ursu, M. 2011. Reasoning for Video-mediated Group Communication. In *ICME*.

Griffin, Z. M., and Bock, K. 2000. What the Eyes Say about Speaking. *Psychological Science (Wiley-Blackwell) 11*, 4, 274.

Jansen, J., Cesar, P., Bulterman, D. C. A., Stevens, T., Kegel, I., and Issing, J. 2011. Enabling Composition-Based Video-Conferencing for the Home. *Multimedia IEEE Transactions on 13*, 5.

Jokinen, K., Nishida, M., and Yamamoto, S. 2009. Eye-gaze experiments for conversation monitoring. In *IUCS '09*, ACM, New York, 303–308.

Jokinen, K., Harada, K., Nishida, M., and Yamamoto, S. 2010. Turn-Alignment Using Eye-Gaze and Speech in Conversational Interaction. *Information Systems Journal*, September, 2018–2021.

Jokinen, K. 2009. Gaze and Gesture Activity in Communication. In *UAHCI '09*, vol. 5615. Springer Berlin / Heidelberg, 537–546.

Rayner, K. 1998. Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin 124*, 3, 372–422.

Richardson, D. C., Dale, R., and Kirkham, N. Z. 2007. The Art of Conversation Is Coordination. *Psychological Science 18*, 5, 407–13.

Rienks, R., Poppe, R., and Heylen, D. 2010. Differences in head orientation behavior for speakers and listeners: An experiment in a virtual environment. *ACM Trans. Appl. Percept. 7*, 1, 2:1—-2:13.

Salvucci, D. D., and Goldberg, J. H. 2000. Identifying fixations and saccades in eye-tracking protocols. In *ETRA '00*, ACM, New York, NY, USA, 71–78.

Ursu, M., Torres, P., Zsombori, V., Frantzis, M., and Kaiser, R. 2011. Entertaining each other from a Distance: Orchestrating Video Communication and Play. In *Proceedings of the ACM Multimedia Conference*.