Journal of VLSI Signal Processing 49, 317–328, 2007

DOI: 10.1007/s11265-007-0081-6

A Multimodal Approach to Communicative Interactivity Classification

TOMASZ M. RUTKOWSKI
Brain Science Institute, RIKEN, Saitama, Japan

DANILO MANDIC
Department of Electrical and Electronic Engineering, Imperial College of Science, Technology and Medicine, London, UK

ALLAN KARDEC BARROS
Laboratory for Biological Information Processing, Universidade Federal do Maranhão, Maranhão, Brazil

Received: 15 May 2006; Revised: 19 October 2006; Accepted: 2 April 2007

Abstract. The problem of modality detection in so called communicative interactivity is addressed. Multiple audio and video recordings of human communication are analyzed within this framework, based on fusion of the extracted features. At the decision level, support vector machines (SVMs) are utilized to segregate between the communication modalities. The proposed approach is verified through simulations on real world recordings.

Keywords: human communication analysis, data fusion, multimedia information processing, audiovisual data fusion

1. Introduction

Multimodal interfaces are an emerging interdisciplinary discipline which involves different modalities of a generic communication process, such as speech, vision, gestures, and haptic feedback. The main goal is to enable better understanding and hence more convenient, intuitive, and efficient interaction between humans and machines (especially computers). Further requirements are that the users ought to interact with such technology in a natural way, without the need for special skills. Emerging work on communicative activity monitoring addresses the problem of automatic activity evaluation in audio and visual channels for distance learning applications [1]. The approaches, however, are limited in that they focus on separated activity in communicative interaction evaluation only, without considering other aspects of the behavioral interdependence in communication.

In this work, we present an attempt to combine knowledge from human communication theory and signal/image processing in order to provide an intelligent way to evaluate communicative situations among humans. This analysis will be used in later stages for implementation of communicative interaction models in human–machine interfaces. For the purpose of evaluation of multimodal interaction, in order to classify them according to "communicative intelligibility" (a measure of potential affordance [2] [usability] of the analyzed interaction), it is necessary to first identify certain illustrative communicative situations from the recorded multimedia streams. We next provide some theoretical background, which is followed by a proposal of a multidimensional interaction evaluation engine, supported with some experimental results.

2. Multimodal Features

The underlying aim of this study is to identify those audio–visual features of the (human) communication process that can be tracked and which, from an information processing point of view, are sufficient to create and recreate the climate of a meeting ("communicative interactivity"). This communicative interactivity analysis provides a theoretical, computational and implementation related framework in order to characterize the human-like communicative behavior.

The analysis of spoken communication is an already mature field, and following the above arguments, our approach will focused on the dynamical analysis of non-spoken components of the communication. In the proposed model of communicative interactivity, based on interactive (social) features of captured situations, two sensory modalities (visual and auditory) of communicative situations are utilized.

The working hypothesis underlying our approach is therefore that observations of the non-verbal communication dynamics contain sufficient information to allow us to estimate the climate of a situation, that is, the communicative interactivity. To that end, the multimodal information about the communication environment, must be first separated into the *communication-related* and *environmental* components[1].' In this way, the audio and video streams can be separated into the information of interest and background noise [3–6].

Highly visually or auditory intensive environments affect the overall impression of the observed/perceived communication situation. On the other hand, since the communicators have usually a limited ability to change the environmental features (i.e. the level of external audio or video), this study recognizes the environmental characteristics as a distinct feature set taken for the overall evaluation of communication. Physical features of the environment can be extracted after separation of the recorded information streams into two categories: items related to the communication process and items unrelated to (useless for) it [4, 6]. The general idea is to split the audio and video streams into background noise and useful signals produced by the communicators. This concept is depicted in Fig. 1, where two sets of cameras capture visual activities and two sets of microphones capture auditory streams in the environment where communicating people are located. The ongoing estimation of mutual information streams in a real conversation is shown in Fig. 2, where two peoples' face-to-face conversation was analyzed. Further signal processing procedures are described in next sections.

We first detect the presence of auditory and visual events that occur in the space but are not related to the communicators' actions (i.e. background audio and video). In the current approach, the analysis of the environmental dimension is performed in two stages:

- Noise and non-speech power level difference extraction
- Non-communication-related visual activity (background flow) estimation

These procedures both can avert the attention of the listeners.

3. Evaluation of Communicative Interactivity

The proper classification of the role in communication members (senders, receivers, transient stages) during the ongoing meeting might be performed with evaluation of audiovisual synchrony, which is a novel idea comparing to existing facial gesture recognition and tracking techniques. The features extracted from recorded speech or nonverbal auditory responses should be synchronized with motion of faces of the communicators. Since recorded audio channels carry too much redundant information, we decided to perform the feature extraction combined with compression. Since the communication act performed by humans incorporates in most situations speech, we decided to use the most common speech features from speech recognition research. We use only first 24 MFCC coefficients obtained as in [7], which carry the significant information from speech occurrences. In case of video we have to obtain the features that would be compatible in dimension to the above audio representations and that would carry information about facial motion.

We desire to obtain video features that carry information about the communication-related motion, and are also compatible with the audio features. Two modalities: The search for faces and moving contours are combined to detect communicating humans in video. This is achieved as follows: For two consecutive video frames $\mathbf{f}_{[h\times w]}(t-1)$ and $\mathbf{f}_{[h\times w]}(t)$, the temporal gradient is expressed as a smoothed difference between the images convoluted with a two-dimensional Gaussian filter $\mathbf{g}$ with the adjusted standard

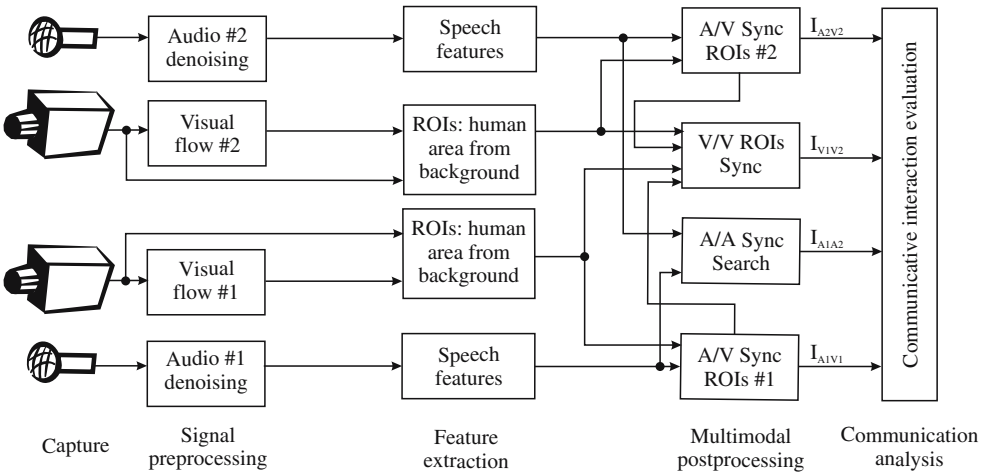

Figure 1. The communicative interaction analysis system chart. Two sets of cameras and microphones capture visual flow information and denoised speech, which is later processed to asses levels of communication interactivity among talking people.

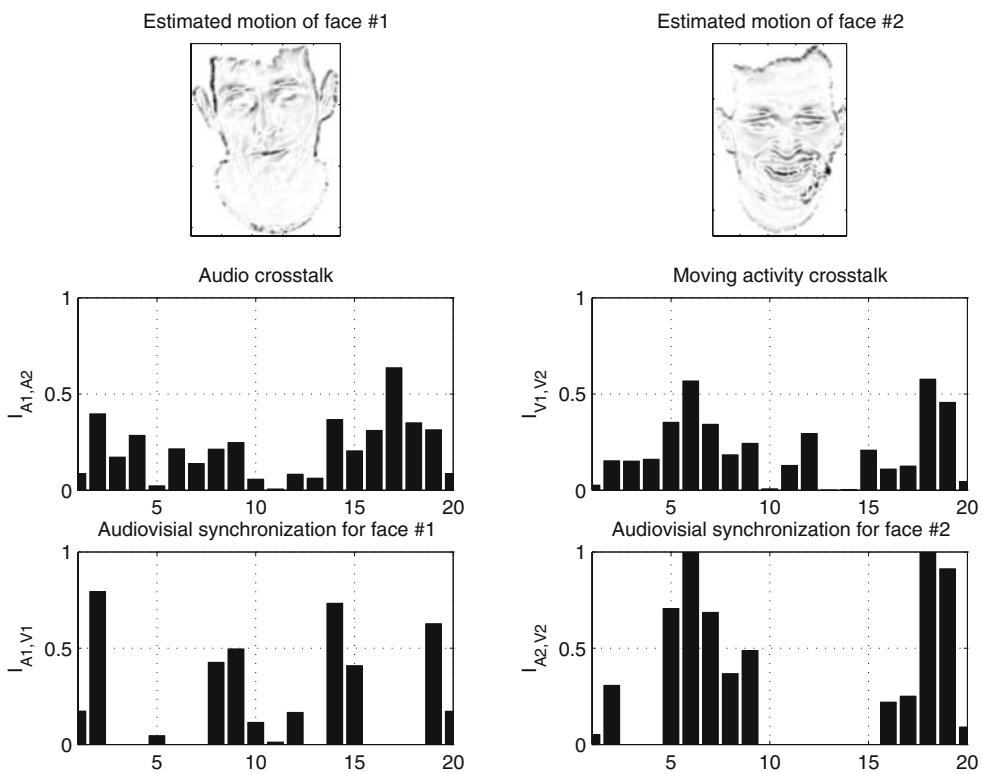

Figure 2. The application to real human conversation. The *two top plots* show the motion features extracted from two facial traced videos. The audiovisual features responsible for local speech to face motion features of every person are plotted as $I_{A1,V1}$ and $I_{A2,V2}$. Plots $I_{A1,A2}$ and $I_{V1,V2}$ showing cross audio–audio and video–video synchronization features present very low activities and small variance, which could be interpreted as a sign for smooth and not disturbed communication event.

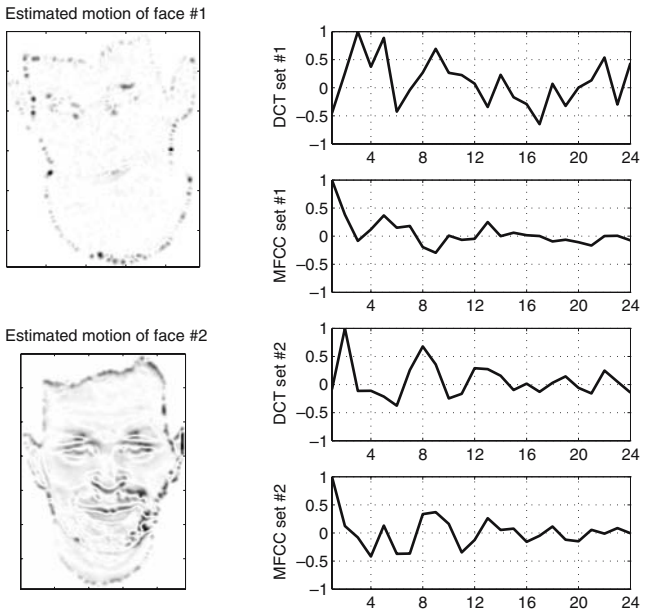

Figure 3. Exemplary figure showing both audio MFCC features and video motion DCT features extracted from both media recorded during the face-to-face communication. Face #1 presents limited activity and both audiovisual features are slightly less dynamic (lower frequency) comparing to face #2.

deviation σ. The pixel $G_{[h\times w]}(t, n,m)$ of the gradient matrix at time t is calculated as follows:

$$\mathbf{G}_{[h\times w]}(t,n,m) = \left|\sum_{i=1}^{x}\sum_{j=1}^{y}\mathbf{d}_{[h\times w]}(t,n-i,m-j)\mathbf{g}_{[x\times y]}(\sigma,i,j)\right|, \tag{1}$$

where $\mathbf{d}_{[h\times w]}(t)$ is the difference between consecutive frames of the size $h \times w$ pixels:

$$\mathbf{d}_{[h\times w]}(t) = \mathbf{f}_{[h\times w]}(t) - \mathbf{f}_{[h\times w]}(t-1). \tag{2}$$

The absolute value is taken to remove the gradient directional information and to enhance the movement capture.

For the face detection and later tracking from the estimated motion information we use the modification of "eigenfaces" features obtained from nonnegative matrix factorization method (NMF) [8]. The modification of NMF in our approach was based on choice of input features which were contours of faces obtained from differential frames of captured video. Features obtained in such a way are more localized and correspond to the intuitive parts of the faces (contours of the face, eyes, nose and mouth). Since we obtain the features from gradient differential images, such method is more suitable to classify the regions as faces and non-faces. To extract only the features with highest energy from motion frames and compress it, making compatible with audio features, we perform the two dimensional digital cosine transformation (DCT). The discrete cosine transform is closely related to the discrete Fourier transform. It is a separable, linear transformation; that is, the two-dimensional transform is equivalent to a one-dimensional DCT performed along a single dimension followed by a one-dimensional DCT in the other dimension. The definition of the two-dimensional DCT for an input image $\mathbf{G}$ and output image $\mathbf{V}$ is:

$$\mathbf{V}(t,p,q) = \sum_{m=0}^{M-1}\sum_{n=0}^{N-1}\mathbf{G}(t,m,n)\cos\frac{\pi(2m+1)p}{2M}\cos\frac{\pi(2n+1)q}{2N} \tag{3}$$
$$0 \le p \le M-1$$
$$0 \le p \le N-1,$$

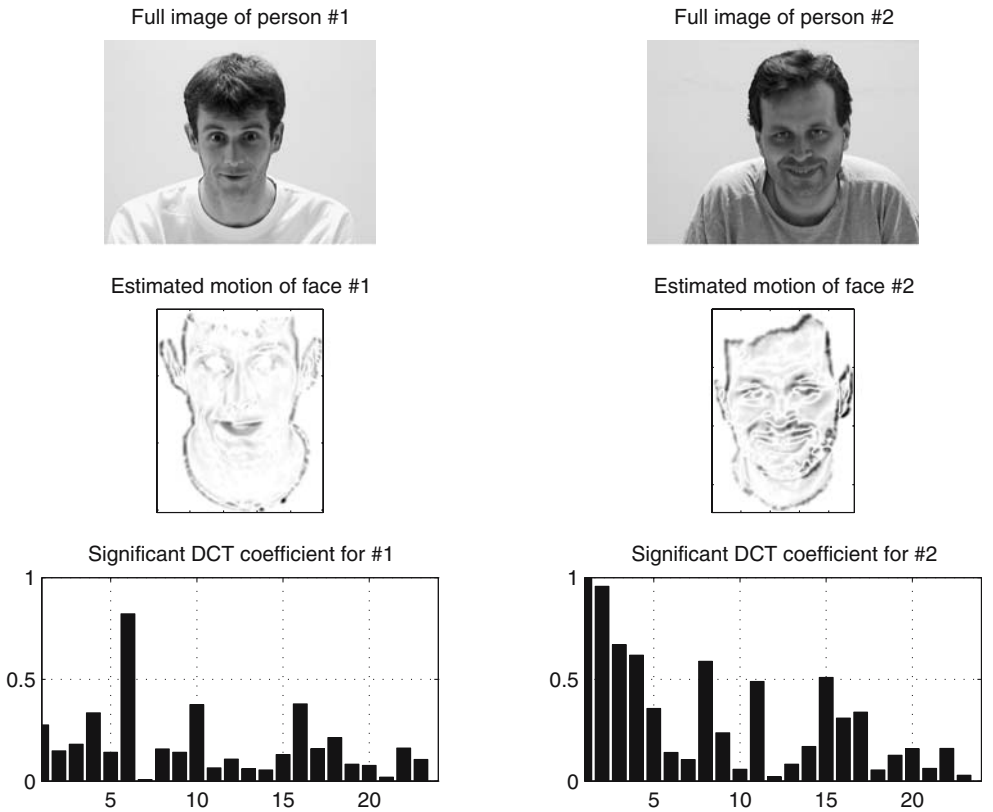

Figure 4. The DCT features extracted from communicating faces. The situation is presented, where both speakers are active in visual mode. In this case DCT features reflecting the activity have comparable amplitudes, but still it is possible to recognize which face is more active for every time slot.

where:

$$\alpha_{p,q} = \begin{cases} 1/\sqrt{M}, \mathrm{p}, \mathrm{q} = 0 \\ \sqrt{2/M}, 1 \leq p, q \leq M-1 \end{cases} \quad (4)$$

Again only 24 first and significant DCT coefficients are taken to be compatible in size with audio MFCC features. The audio and video features obtained from active and nonactive member of dyadic communication situation are presented on Fig. 3. The examples of DCT features only are shown on Fig. 4 presents the situation where both speakers are active in visual mode in interlaced fashion. In this case DCT features reflecting the activity have similar amplitude, but

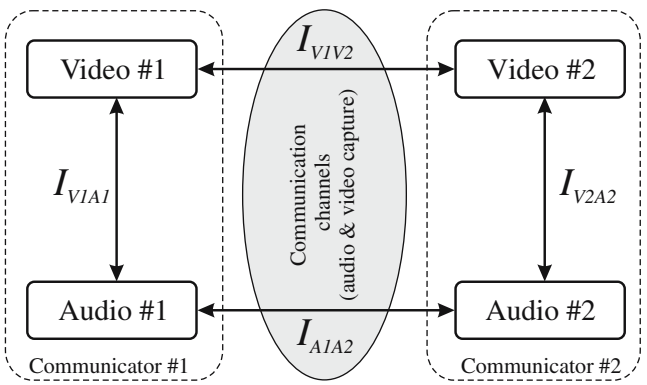

Figure 5. Scheme for the communicative interactivity evaluation. Mutual information estimates $I_{A_1V_1}$ and $I_{A_2V_2}$ between audio and visual features streams of localized communicators account for the local synchronization. The estimates $I_{A_1A_2}$ and $I_{V_1V_2}$ are to detect cross talks in the same modality.

still it is possible to recognize which face is more active at every time slot. If we consider the above features sets as being independent samples from a multivariate probability distribution $p(A,V)$, then the proper measure for audiovisual synchrony or asynchrony is mutual information $I_{A,V}$ between random variables A and V. Since the distributional forms of $p(A)$ and $p(V)$, $p(A,V)$ are unknown the assumption of continuous distribution can be made. The features vectors MFCC and DCT are considered in such a case as samples locally Gaussian, multivariate distribution $p(A,V)$. Presence of communication is judged based on mutual information(s) between visual and audio features for selected regions of interest (ROI) [9], as:

$$\begin{aligned} I_{A_iV_i} &= H(A_i) + H(V_i) - H(A_i,V_i) \\ &= \frac{1}{2}\log(2\pi e)^n|R_{A_i}| + \frac{1}{2}\log(2\pi e)^m|R_{V_i}| - \frac{1}{2}\log(2\pi e)^{n+m}|\mathbf{R}_{A_iV_i}| \\ &= \frac{1}{2}\log\frac{|R_{A_i}||R_{V_i}|}{|R_{A_iV_i}|}, \end{aligned} \tag{5}$$

where $i = 1,2$ and R_{A_i}, R_{V_i}, $R_{A_iV_i}$ stand for empirical estimates of the corresponding covariance matrices of the feature vectors [10] (computed recursively).

Simultaneous activity estimates in the same modes (audio and video, respectively) are calculated for video and audio streams, respectively, as:

$$\begin{aligned} I_{V_1V_2} &= \frac{1}{2}\log\frac{|R_{V_1}||R_{V_2}|}{|R_{V_1V_2}|} \quad \text{and} \\ I_{A_1A_2} &= \frac{1}{2}\log\frac{|R_{A_1}||R_{A_2}|}{|R_{A_1A_2}|}, \end{aligned} \tag{6}$$

where $R_{A_1A_2}$ and $R_{V_1V_2}$ are the empirical estimates of the corresponding covariance matrices for unimodal feature sets representing different communicator activities. A_1, A_2 and V_1, V_2 are audio and video features extracted from communicator #1 and #2, respectively. Quantities $I_{A_1V_1}$ and $I_{A_2V_2}$ evaluate the local synchronicity between the audio (speech) and visual (mostly facial movements) flows and it is expected that the sender should exhibit the higher

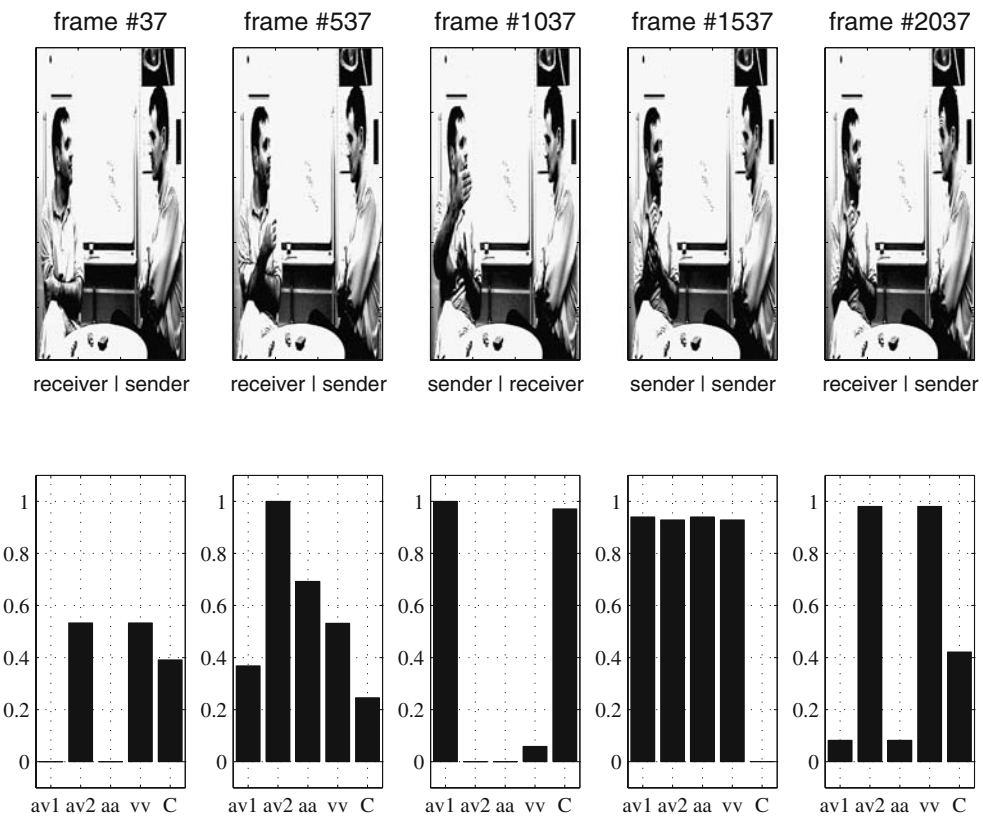

Figure 6. All mutual information tracks together with efficiency estimate (av1; av2; aa; vv; C; stand, respectively, for $I_{A_1}V_1$; $I_{A_2}V_2; I_{A_1}A_2; I_{V_1}V_2; C(t)$;).

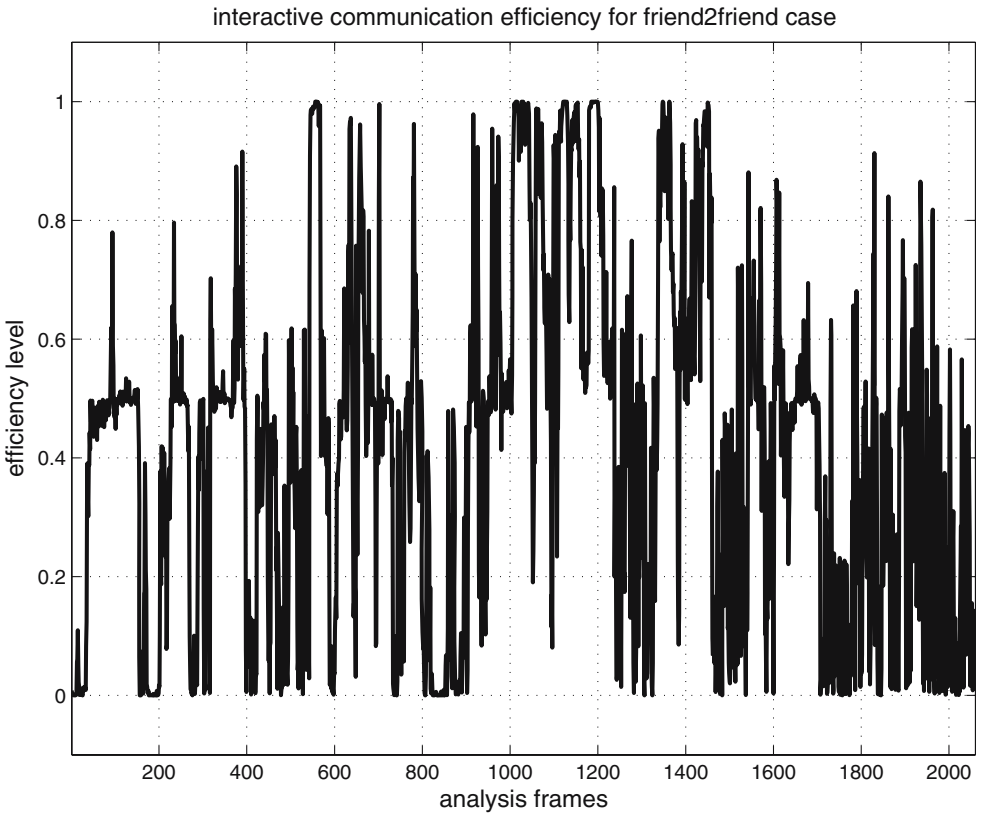

Figure 7. An example of efficiency level $C(t)$ as in Eq. (7) in face-to-face communication. The values vary over time in range from 0 showing non-efficient communicative situation to 1 for perfectly efficient communicative situations.

synchronicity, reflecting the higher activity. Quantities $I_{V_1V_2}$ and $I_{A_1A_2}$ are related to the possible cross talks in same modalities (audio–audio, video–video). The latter is also useful to detect the possible activity overlapping, which can impair the quality of the observed communication.

Communicative interactivity evaluation assesses the behavior of the participants in the communication from the audio–visual channel, and reflects their ability to "properly" interact in the course of conversation. This is quantified by synchronization and interaction measures [4, 6]. In [10] the *communication efficiency* is defined as *a measure that characterizes the behavioral coordination of communicators*. Here, a measure of the communication efficiency is proposed as a combination of four estimates of mutual information [11]: (1) two visual (V_i), (2) two audio (A_i), and (3) two pairs of audiovisual features ($A_i; V_i$). Figure 5 shows the concept of utilization of mentioned in previous section unimodal and multimodal mutual information measure to asses to communication interactivity level. Efficiency a qualitative measure of the communication process directly related to the attention level and to the dynamic involvement of the communicators [12]. A combined measure of temporal communication efficiency can be calculated as:

$$C(t) = \left(1 - \frac{I_{V_1V_2}(t) + I_{A_1A_2}(t)}{2}\right) \times |I_{A_1V_1}(t) - I_{A_2V_2}(t)|, \quad (7)$$

and it allows us to monitor communicative process taking into account all four (in case of face-to-face communication) mutual information estimates. In case of efficient communication time frames it reaches level two, when there are no cross talks in auditory or visual streams, suggesting that only single party is active. In case of transient stages or non-efficient situations (uni- or multimodal cross

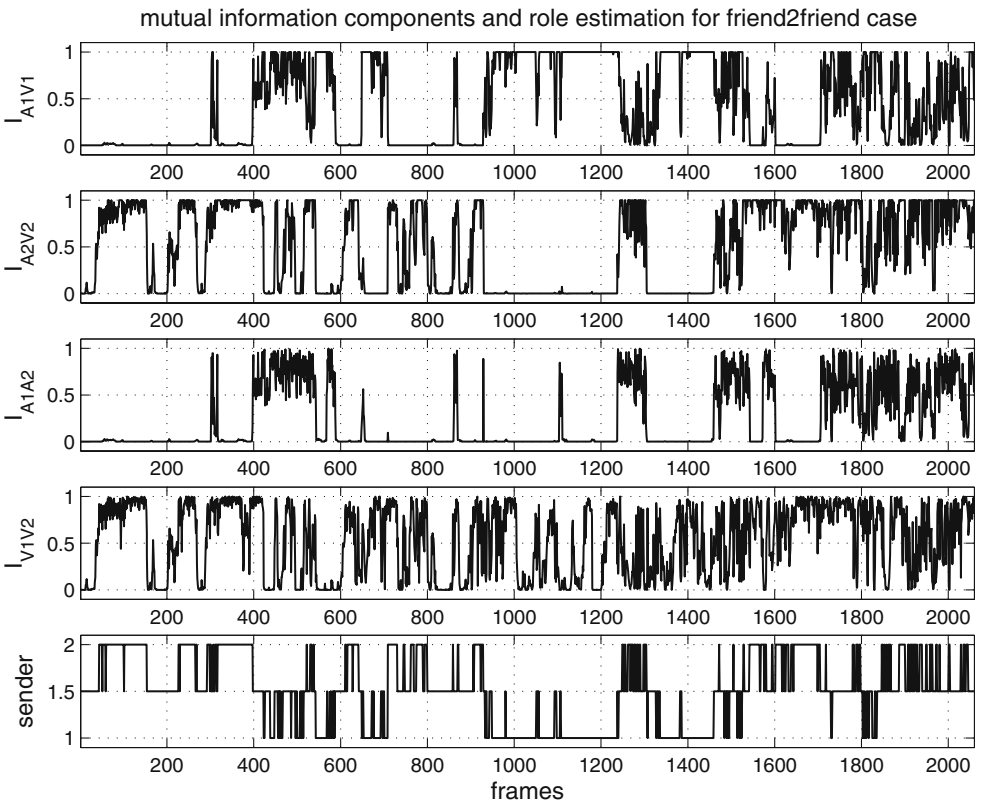

Figure 8. An example showing audiovisual activities and role in conversation estimation. The *four panels* starting from the top show two multimodal and two unimodal mutual information estimates of active face-to-face communication. The *bottom panel* present the situation classification result based on learned SVM classifier as in Eq. (3) using mutual information features plotted in above panels.

talks) its value can be in a range $[0, 1]$. The levels of communication efficiency $C(t)$ for different frames of captured communication together with unimodal and multimodal mutual information estimates is shown in Fig. 6. The track in time of efficiency level $C(t)$ dynamics of ongoing communication is also shown in Fig. 7. The efficiency estimation track shows how dynamic conversation could.

The communicator's role, that is, (*sender* or *receiver*) can be estimated by monitoring the behavior of audiovisual features over time. An indication of higher synchronization across the audio and video features characterizes the active member, the sender, while the lower one indicates the receiver (see bottom plot in Fig. 8). This synchronized audiovisual behavior of the sender and the unsynchronized one of the receiver characterizes an efficient communication [4, 6, 10].

This information is used to classify the role in communication situation (*sender*, *receiver* or *transient*) according to the differences in local audiovisual activities for every time frame. The cross audio–audio and video–video ($I_{A1,A2}$ and $I_{V1,V2}$, respectively) are used for *transient* stages detection. Examples of mutual information levels tracks in active face-to-face communication are shown in Fig. 2. The pair of the mutual information estimates for the local synchronization of the senders and the receivers in Eq. (5) is used to give clues about concurrent individual activities during the communication event, while the unimodal cross-activities estimates in Eq. (6), are used to evaluate the interlaced activities for a further classification. Intuitively, the efficient sender–receiver interaction involves *action* and *feedback*. The interrelation between the actions of a sender and feedback of a receiver is therefore monitored, whereby the audio–visual synchronicity is used to determine the roles.

In our approach, the interactions between individual participants in communication are modeled

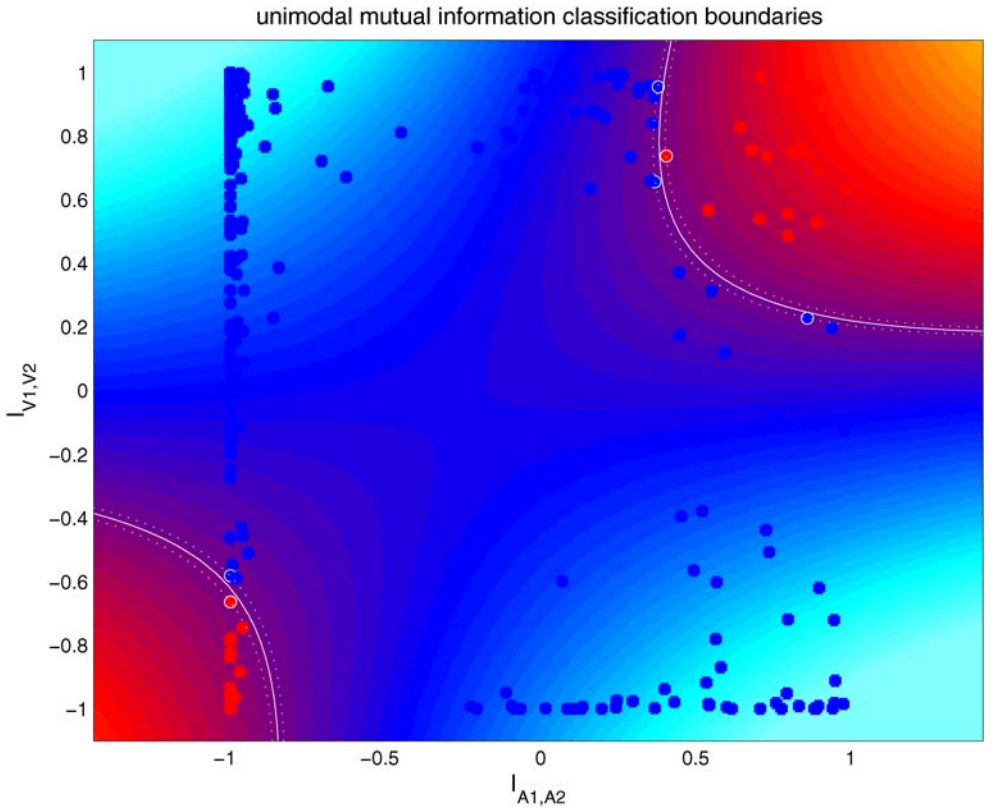

Figure 9. Example of learned decision boundaries of SVM RBF classifier learned using only two features sets $I_{A1,A2}$ versus $I_{V1,V2}$ for later discrimination of different communication interactivity levels.

within the *data fusion* framework, based on features coming simultaneously from both the audio and video. A multistage and a multisensory classification engine [13] based on the support vector machine (SVM) approach is used at the *decision making* level of the data fusion framework, where the *one-versus-rest-fashion* approach is used to identify the phases during ongoing communication (based on the mutual information estimates from Eqs. (5) and (6). The decision is made based highest classification output from binary SVMs.

At the decision level SVMs are particularly suited when *sender–receiver* or *receiver–sender* situations are to be discriminated from the *noncommunicative* or *multi-sender* cases. In this work, a kernel based on a radial basis function (RBF) is utilized [14], and is given by:

$$K(\mathbf{x}, \mathbf{x}_i) = e^{-\gamma||\mathbf{x}-\mathbf{x}_i||^2} \tag{8}$$

where γ is a kernel's width parameter. Using the above concept, an arbitrary multimodal mutual information combination for unimodal cases ($I_{A1,A2}$ and $I_{V1,V2}$) λ can be categorized into four categories:

(1) $f(\lambda) \in (-\infty, -\alpha]$ for the *noncommunicative* case with no interaction (no communication or a single participant)
(2) $f(\lambda) \in (-\alpha, 0]$ for the *sender–receiver* case
(3) $f(\lambda) \in (0, \alpha)$ for the *receiver–sender* case
(4) $f(\lambda) \in [\alpha, +\infty)$ for the *sender–sender* case

The categories (1) and (4) are somehow ambiguous due to the lack of clear separation boundaries, and they are treated by separately trained SVM classifiers. The threshold α usually is set that $\alpha \in [0.4, 0.5]$ [6]. Example of classifiers' decision boundaries are shown for two-dimensional case for the unimodal example in Fig. 9. For the multimodal case ($I_{A1,V1}$ and $I_{A2,V2}$) where the communicative interactivity is mostly evaluated, the classification categories can be designed as:

(1) $f(\lambda) \in \{(-\infty, -\beta] \cap (\beta, +\infty]\}$ for the *efficient* case with good interaction;

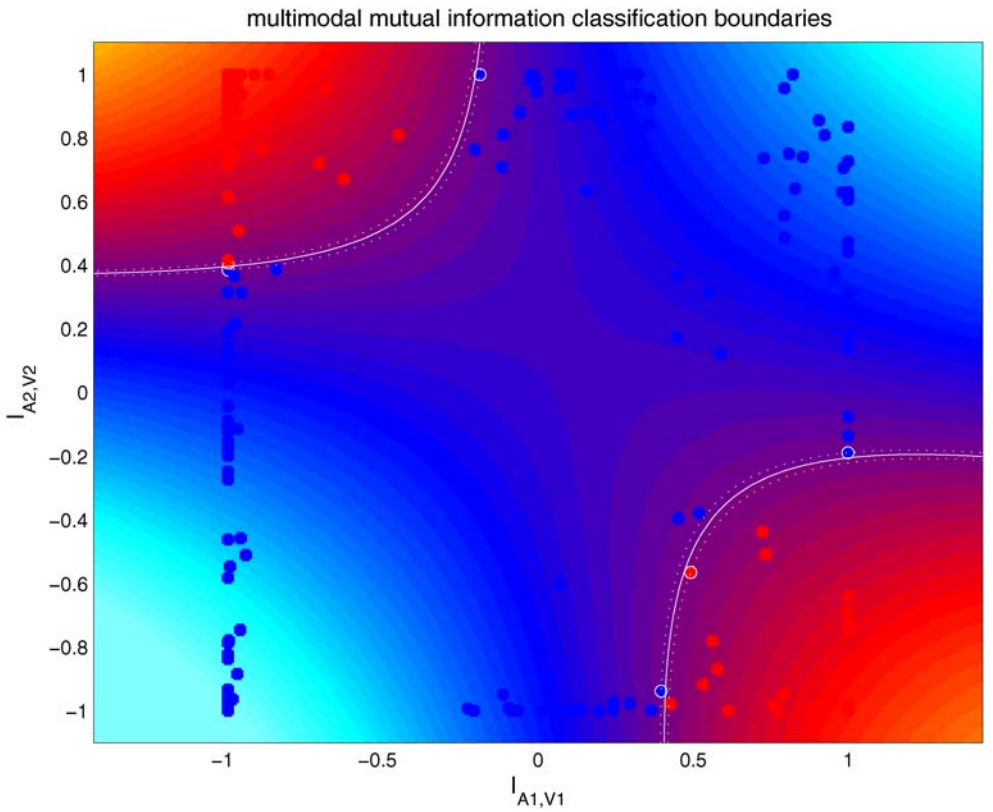

Figure 10. Example of learned decision boundaries of SVM RBF classifier learned using only two features sets $I_{A1,V1}$ versus $I_{A2,V2}$ for later discrimination of different communication interactivity levels.

(2) $f(\lambda) \in (-\beta, \beta)$ for the *inefficient* case when interaction is not proper due to overlapping activities of communicators.

Similarly like for above α threshold, $\beta \in [0.4, 0.5]$. Example of classifiers' decision boundaries for the interactivity classification are shown for two-dimensional case for the multimodal example in Fig. 10.

4. Experimental Results and Conclusions

The experiments, where the participants in communication were engaged in a face-to-face conversation, were conducted to validate the proposed approach. Six videos of ongoing conversation were shown with three different setups: two of teacher–student conversations; two of talking friends; and two of coworkers professional discussions. In the presented

Table 1. Comparison of objective and subjective (seven experts) communication interactivity evaluations (the score around 100% would suggest fully interactive event, while lower one characterizes overlapped discourse between communicators).

Case	Objective (proposed method; %)	Subjective (human experts; %)	Method's error (%)
Teacher and student #1	51	50	1
Teacher and student #2	63	60	3
Friends #1	56	70	−14
Friends #2	75	80	−5
Coworkers #1	63	75	−12
Coworkers #2	60	70	−10

approach we did not take into account any social relations of communicating people, so the different chosen examples should reflect only different dynamics of ongoing processes.

In order to evaluate proposed approach we compared its results w subjective evaluation of "human experts" which were asked to watch and evaluate ongoing communication videos based on subjective estimation of amount of overlapping in conversation for every role change in sender–receiver roles. Also the overall impression interactivity was taken into account. Results of seven experts decisions are shown in last third column in Table 1. The results of proposed approach evaluate the communication interactivity level showing similar performance to that of subjective evaluations of human experts for the analyzed videos is summarized second column in Table 1.

This way, the proposed *data fusion* approach for the evaluation of communicative interaction represent a step forward in the modeling of communication situation, as compared to the existing audio- and video-only approaches [1]. The experiments have clearly shown the possibility to estimate the interactivity level, based on the behavioral analysis of the participants in communication. The mutual information based feature extraction of multimodal audio and video data streams makes it possible to detect the presence participants and to classify them according their role. Despite some difference between the conclusions of a seven human experts and the proposed method, our results show strong correlation between the two. In fact, the human judgement is also highly subjective, therefore further studies will have a larger population of human experts to balance their opinion.

Acknowledgements

Authors would like to thank Prof. Toyoaki Nishida, Prof. Michihiko Minoh, and Prof. Koh Kakusho of Kyoto University for their support and fruitful discussions in frame of the project "Intelligent Media Technology for Supporting Natural Communication between People", which was partially supported by the Ministry of Education, Science, Sports and Culture in Japan, Grant-in-Aid for Creative Scientific Research, 13GS0003, where presented approach was developed. Also we would like to thank for many discussions to Prof. Victor V. Kryssanov of Ritsumeikan University in Kyoto at beginning stages of presented research, which were very valuable to shape the final approach.

Note

1. Notice the analogy to the Wold decomposition theorem which states that every signal can be decomposed into its deterministic and stochastic part.

References

1. M. Chen, "Visualizing the Pulse of a Classroom," in *Proceedings of the Eleventh ACM International Conference on Multimedia*, ACM Press, 2003, pp. 555–561.
2. J.J. Gibson, "The Theory of Affordances," in *Perceiving, Acting and Knowing*, R. Shaw and J. Bransford (Eds.), Erlbaum, Hillsdale, NJ, 1977.
3. T.M. Rutkowski, M. Yokoo, D. Mandic, K. Yagi, Y. Kameda, K. Kakusho and M. Minoh, "Identification and Tracking of Active Speaker's Position in Noisy Environments," in *Proceedings of International Workshop on Acoustic Echo and Noise Control (IWAENC2003)*, Kyoto, Japan, 2003, pp. 283–286.
4. T.M. Rutkowski, K. Kakusho, V.V. Kryssanov and M. Minoh, "Evaluation of the communication atmosphere," *Lect. Notes Comput. Sci.*, vol. 3213, 2004, pp. 364–370.
5. T.M. Rutkowski, Y. Yamakata, K. Kakusho and M. Minoh, "Smart sensor mesh—intelligent sensor clusters configuration based on communicative affordances principle," *Lecture Notes in Artificial Intelligence*, vol. 3490, 2005, pp. 147–157.
6. T.M. Rutkowski and D. Mandic, "Communicative interactivity—a multimodal communicative situation classification approach," *Lect. Notes Comput. Sci.*, vol. 3697, 2005, pp. 741–746.
7. S. Furui, "*Digital Speech Processing, Synthesis, and Recognition—Second Edition, Revised and Expanded. 2nd edn. Signal Processing and Communications Series*," Marcell Dekker, Inc., New York, Basel, 2001.
8. D.D. Lee and H.S. Seung, "Learning the Parts of Objects by Non-Negative Matrix Factorization," *Nature*, vol. 401, 1999, pp. 788–791.
9. A. Hyvarinen, J. Karhunen, E. Oja, "*Independent Component Analysis*," Wiley, 2001.
10. T.M., Rutkowski, S. Seki, Y. Yamakata, K. Kakusho and M. Minoh, "Toward the Human Communication Efficiency Monitoring from Captured Audio and Video Media in Real Environments," *Lect. Notes Comput. Sci.*, vol. 2774, 2003, pp. 1093–1100.
11. C. Shannon and W. Weaver, "*The Mathematical Theory of Communication*," University of Illinois Press, Urbana, 1949.
12. V. Kryssanov and K. Kakusho, "From Semiotics of Hypermedia to Physics of Semiosis: A view from System Theory," *Semiotica*, vol. 154, no. 1/4, 2005, pp. 11–38.
13. C.W. Hsu and C.J. Lin, "A Comparison of Methods for Multi-Class Support Vector Machines," *IEEE Trans. Neural Netw.*, vol. 13, 2002, pp. 415–425.
14. V. Cherkassky and F. Mulier, "*Learning from Data. Adaptive and Learning Systems for Signal Processing, Communication, and Control*," Wiley, USA (1998).

Dr. Rutkowski received his Ph.D. in Technology (telecommunications and acoustics) in 2002 from Wroclaw University of Technology, Wroclaw, Poland. He completed a postdoctoral training in multimedia in Department of Intelligence Science and Technology at Kyoto University, Japan. He is currently a Research Scientist in Brain Science Institute RIKEN, Japan. He is currently involved in research on brain computer/machine interface, which covers problems of neuroscience and multimedia. He has written about 80 publications in multimedia, brain and telecommunications signal processing/modeling research fields.

Dr. Mandic received his Ph.D. degree in nonlinear adaptive signal processing in 1999 from Imperial College, London, London, U.K. He is now a Reader with the Department of Electrical and Electronic Engineering, Imperial College London, London, U.K. He has written about 200 publications on a variety of aspects of signal processing, a research monograph on recurrent neural networks and has coedited a book on signal processing for information fusion. He has been a Guest Professor at the Catholic University Leuven, Leuven, Belgium and Tokyo University of Agriculture and Technology (TUAT), and Frontier Researcher at the Brain Science Institute RIKEN, Tokyo, Japan. Dr. Mandic has been a Member of the IEEE Signal Processing Society Technical Committee on Machine Learning for Signal Processing, Associate Editor for IEEE Transactions on Circuits and Systems II, Associate Editor for International Journal of Mathematical Modeling and Algorithms, and Associate Editor for the IEEE Transactions on Signal Processing. He has won awards for his papers and for the products coming from his collaboration with industry.

Dr. Barros received his B.S. degree in Electrical Engineering from Universidade Federal do Maranhao, Brazil in 1991, his M.S. degree in Information Engineering from Toyohashi University of Technology, Toyohashi, Japan in 1995, and his D.Eng. degree from Nagoya University, Nagoya, Japan in 1998. He worked as a Frontier Researcher from 1998 to 2000 at The Institute of Physical and Chemical Research (RIKEN), Japan. Currently, he is an Associate Professor at Universidade Federal do Maranhao, Brazil. His research interests include biomedical engineering, information coding, and speech signal processing.