# Embodied Communicative Activity in Cooperative Conversational Interactions - Studies in Visual Interaction Management

**Kristiina Jokinen [1], Stefan Scherer [2]**

[1] Institute of Behavioural Sciences
University of Helsinki
PO Box 9, Helsinki
FIN-00014, Finland
E-mail: kristiina.jokinen@helsinki.fi

[2] Speech Communication Laboratory
Centre for Language and Communication Studies
Trinity College Dublin
7-9 South Leinster Street
Dublin 2, Ireland
E-mail: stefan.scherer@gmail.com

*Abstract: Non-verbal communication is important in order to maintain fluency of communication. Gestures, facial expressions, and eye-gazing function as efficient means to convey feedback and provide subtle cues to control and organise conversations. In this article, we discuss the relation of verbal and non-verbal feedback from the point of view of communicative activity, and focus especially on hand gestures and body movement in the coordination of the interaction. This is called "Visual Interaction Management". Combining the top-down approach, i.e. manual annotation and analysis of the data, with the bottom-up analysis of the speech and visual signals, we can visualize the speakers' speech and gesture activity, and align this with those gestures and body movements that the interlocutors interpret as communicatively important. As the method for the bottom-up analysis, we use Echo State Networks, an architecture for recurrent neural networks, and use it in the recognition analysis of communicative behaviour patterns.*

*Keywords: Embodied Communicative Activity; Naturalistic Interaction; Interaction Management*

# 1 Introduction

Social communication refers to interactions where the participants, besides trying to achieve a possible task goal, also intend to maintain mutual relations and a good atmosphere, and the interaction as a whole would appear cooperative and pleasant. In these interactions natural language communication has an important role, but also various non-verbal means are effectively used to construct shared knowledge and to create social bonds.

In this article we will explore especially the connection of spoken language and gesturing, i.e. hand gestures and body movements. Gesturing has an iconic function to describe events for the partner, but they also contribute to the needs of communication management and directing the partners' attention. The "meaning" of the gestures thus correlates with the content on the utterance level, but also gets interpreted in a larger context of the communicative needs in the dialogue situation and activity in which the participants are involved. Ultimately, the meaning includes the whole cultural context of the interlocutors.

As evidenced by previous research, verbal and non-verbal communication are simultaneously produced [25], and thus there is a link between communicatively relevant body and hand movement and the coordination and control of interactive situations. There are also views which consider language, or its evolution, intimately linked to gestures, and the view has become popular in the past ten years, especially in neuro-scientific and cognitive studies [12].

In this article we focus on gesturing and body movement, and provide preliminary analyses of the use and function of these activities in dialogues. We will study especially the regulating and coordinating function of gesticulation in communicative situations, and thus we refer to this kind of conversation as Visual Interaction Management. We will also visualise the interlocutors' alignment as part of the dialogue activity and discuss how the different non-verbal signals correlate with the verbally expressed content.

Our methodology combines the top-down approach of linguistically annotated corpus into a signal-level analysis of the same video corpus. We can thus coordinate the dialogue activity from the two view-points, which allows us to visualise how the higher-level conceptual labels coincide with signal-level observations of the conversational activity. With the help of the visualisations we can distinguish the interlocutor's signal-level behaviour, and associate this with communicatively meaningful interpretations of particular gestures, body postures, and facial expressions. The specific question is if we can notice that these signs do not only accompany or complement the spoken content, but also can function as independent means for communication management.

For the signal-level analysis, we use the Echo State Networks, an architecture for recurrent neural networks, which are capable of pattern recognition and production. A trained network can also be reversed and used to generate behavioural patterns.

We explore the connection between the patterns and the features so as to evaluate if it could be used to generate behaviour in an agent.

The article is structured as follows. Section 2 discusses related work as background for our studies, and Section 3 presents two examples from the corpus as representatives of cooperative visual interaction management. Section 4 describes the Echo State networks and their use in the experiments. Section 5 shows visualisation of the dialogue activity and discusses Visual Interaction Management. Section 5 presents conclusions and points to further research topics are.

# 2 Related Work

Previous research has already established several ways in which different interaction modalities contribute to smooth communication in the form of rich, articulated feedback [14]. For instance, the use of prosodic and syntactic features in turn-taking and backchannelling has been widely studied (e.g. [27]), as have various types of non-verbal feedback [40] as well as the use of prosodic cues in discourse structuring (e.g. [9]). Also eye-gaze has been established as an important indication of turn-taking [4], as well as signalling the focus of shared attention in meetings [32] and in communication in general [17, 33, 31, 21]: gaze direction serves to frame the interaction and establish who is going to speak to whom. Facial signs in interaction also play an important role [16]. Also the use of gestures in interaction is also well studied [25, 5]. Simulations of conversational behaviour are presented e.g. by [32], who focus especially on turn-taking and feedback, and gesture generation for animated agents and virtual humans has been discussed e.g. in [3, 10, 36, 26]. The combination of gestures with emotional expressions is discussed in [30], and their use with robotic companions which can engage into interaction with the user using a whole range of modalities is exemplified by [6]. Also, cultural differences need to be paid attention to, and work on this line can be found e.g. in [22, 35]. Recently non-verbal communication has attracted interest also in second language learning and behavioural studies: e.g. teachers can observe learners' gestures to assess the proficiency and progress of grammar [15].

Kendon [25] defines gestures as the spontaneous movement of hands and arms, and specifies communicative gestures as those which are intended to be interpreted in the communicative context as carrying some meaning. Not all movements are relevant in communication nor intended to carry a meaning, so there is a continuum from communicative to general movements. For instance, coffee-drinking usually is a non-communicative action, although it could also be used to signal the person is not engaged in the conversation. There are highly conventionalised gestures such as emblems (culturally conditioned signals) and sign languages (a whole framework of conventionalised gestural signs). Kendon [25] has identified different gesture families, which refer to a group of gestures with a similar shape and a semantic theme of their own. For instance, the Open Hand Supine ("palm up") families express gen-

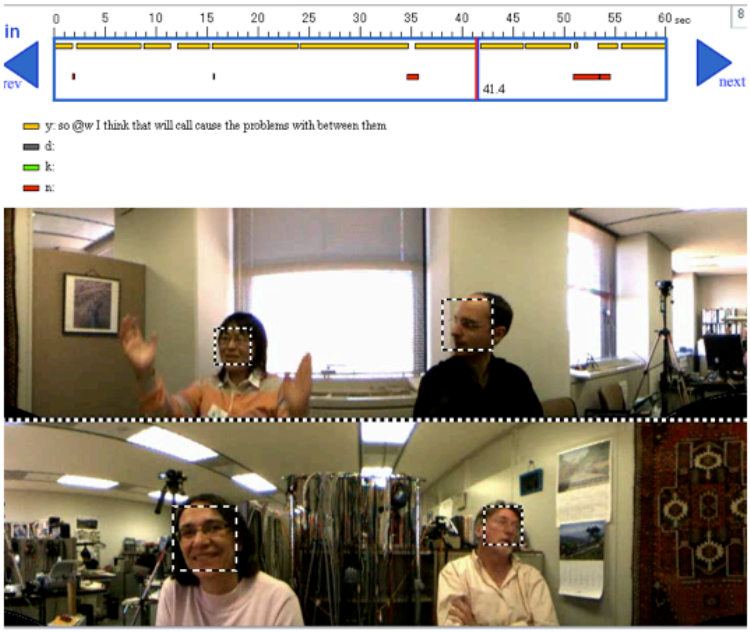


Figure 1
Gesture expressing "will cause problems between them"

eral offering and giving of ideas, whereas the subgroup Open Hand Supine Vertical (movement up-down) is used in relation to cutting, limiting, or structuring information (see example in Figure 1 below). The Index Finger Extended is another gesture family, with the main semantic theme similar to that of the Open Hand family except that it correlates with precise and specific ideas (cf. Figures 3, 4). Some gestures also function on the conversational metalevel [25, 21], and have also been called stand-up gestures by [23].

# 3 Cooperative Conversational Interactions

To investigate the interlocutor's non-verbal activity, we use the video corpus collected in an international setting at the ATR Research Labs in Japan [7]. The corpus consists of three approximately 1.5-hour long, casual conversations between four participants, and they were recorded during three consecutive days. The technical setup included a 360 degree camera and one microphone and is similar to [13]. Conversational topics were not restricted and the speakers did not have any particular task to work on. No specific instructions were given to the speakers either, and chatty interactions quickly got going and proceeded in a cooperative way. The inter-

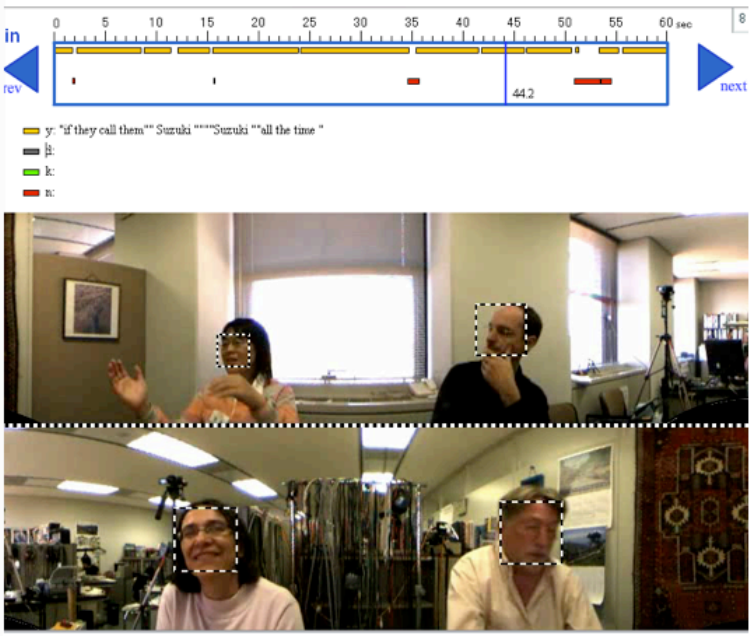


Figure 2
Gesture expressing “if they call them Suzuki, Suzuki all the time”

locutors spoke English but represented different cultural backgrounds and language skills. One of them knew all the other participants while the others were unfamiliar with each other. All shared some background knowledge of the culture and living in Japan.

The data is transcribed, and part of it is annotated with respect to non-verbal communication using the MUMIN coding scheme [2]. Gestures, facial expressions, and body posture are annotated with respect to their form and function. The form features for gestures include the shape of hand, palm, fingers, and hand movement, the features for face and head include the shape and combination of eyes, eye-brows, mouth, and the movement of head, and features for body posture include leaning backward and forward. Each communicative event is also interpreted with respect to turn-taking and feedback giving functions, and annotations also take into account the general semiotic meaning of communicative elements: they can be indexical (pointing), iconic (describing), and symbolic (conventional) signs.

## 3.1 Iconic Gestures and Description of Information

Figure 1 depicts a situation in the beginning of the analysed clip where the speaker (upper panel on left) is explaining how problems may arise between two groups

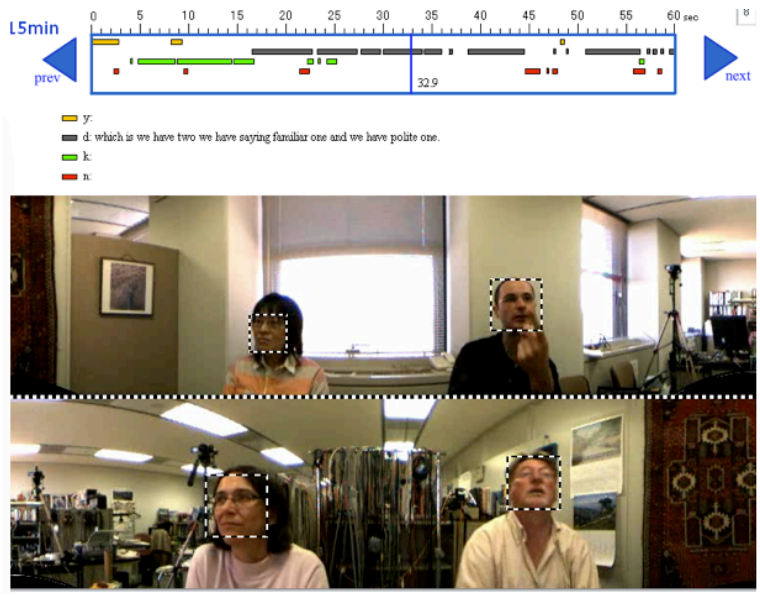


Figure 3
Gesture and facial expression when the speaker explains "two ways of saying a familiar one and a polite one"

of people. She had been talking about how she went to the airport and saw some Japanese business men greeting their foreign visitors and introducing themselves saying "My name is Fujita, I'm Suzuki", and she was worried that misunderstandings may arise if the visitors are not familiar with the Japanese name conventions and would call their hosts without the suffix "san" or "kun". The speaker's gesticulation is large and vivid, and her many iconic gestures paint the scene in front of the listeners: e.g. greetings ("How are you") are illustrated by hand shakes, and "overseas guests" are shown to come from far away by an extended arm and "placing" them in a particular location in the scene, while the speaker's own worry is placed on herself by drawing the arms towards her head and rolling the hands around. The hand shape in the gestures is usually round, with the fingers loosely curled towards the palm and the index finger slightly extended. However, in Figure 1, the hands are clearly open and straight, and their movement is up and down, as if emphasizing the two groups and the problem "between them".

Figure 2 vividly depicts the situation where the guests (originally placed on the speaker's left) call their hosts (on the speaker's right) "Suzuki, Suzuki, all the time, like calling John or Bill", and the speaker's left hand makes swiping movements rhythmically towards the right hand, as if calling someone to catch their attention.

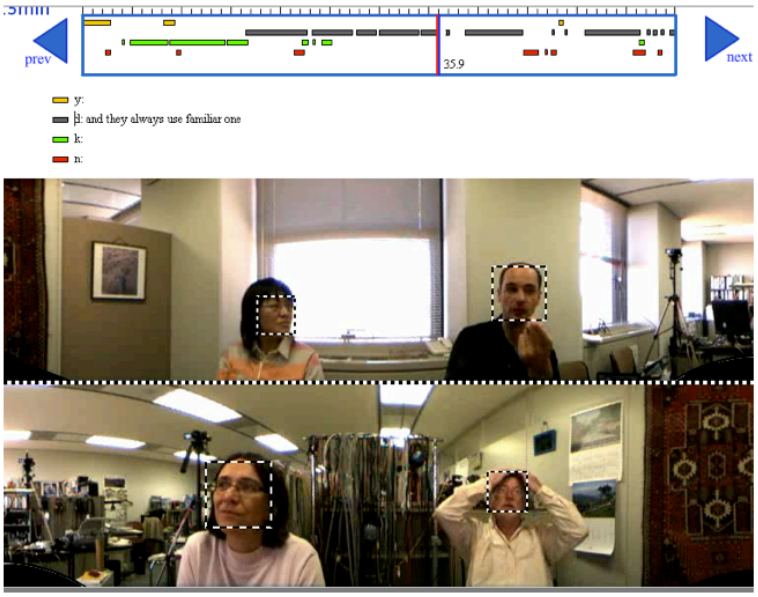


Figure 4
Gesture and facial expression when the speaker explains “they always use the familiar one”. A partner has withdrawn in the background

## 3.2 Indexical Gestures and Structuring of Dialogue

In Figures 3, 4, and 5, the previous dialogue continues, but the topic deals with the distinction in French between tu and vous. The speaker (upper panel right) explains that French has two pronouns, a familiar one and a polite one, and that the use of tu would in some cases be extremely rude, but if one can hear from the accent that the person is not French, then it’s ok. The speaker effectively uses his left hand to emphasise and structure his speech. In Figure 3 the two extended fingers refer to the two different pronouns, whereas in Figure 4, the extended index finger singles out the “familiar one”. In Figure 5, the whole hand refers to a person speaking with a foreign accent, and simultaneously also functions as a turn holder. The speaker also looks at the partners and turns his face frequently between them so as to elicit understanding from them.

In this clip we also see the non-speaking participant’s body movement, accompanied by the large movements of his arms, to signal his participation in the conversation. The partner (lower right) goes backwards when he withdraws himself from the centre as if to ponder upon the presented information (Figure 4), and then comes back to the conversation with a clarification question “and vous is rude?” (Figure 5), leaning forward to show interest and to control the information flow.

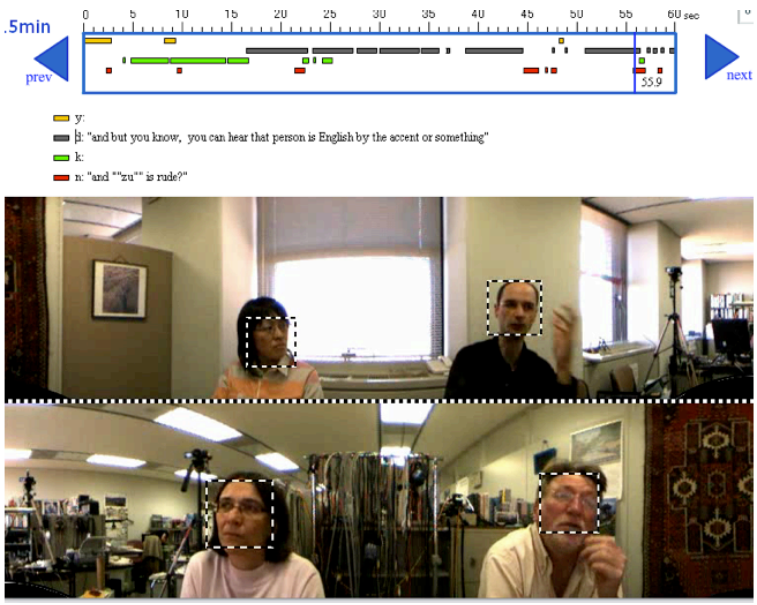


Figure 5
Gesture and facial expression when the speaker utters “you can hear the person is English by accent or something”. The partner simultaneously leans forward and asks a clarification "and vous is rude?”

## 3.3 Visual Interaction Management

In cognitive linguistics studies, the speaker and the hearer are regarded as cooperating agents and their communication is seen as an instance of alignment [34]. For instance, [1] talks about Own Communication Management (OCM) and Interaction Management (IM), referring to the aspects of communication that concern meta-level control of the interaction, such as repairs, initiations of topics, direction of the focus of attention etc. OCM refers to the agent monitoring her own production while IM includes the interaction situation as whole and mainly concerns the agent’s intentions to influence the flow of interaction. Alignment can be exemplified in different language contexts through intercultural studies, and important research questions in this respect have dealt with how to react to the partner’s contributions, how to refer to particular objects in the environment, and how to construct a shared ground (e.g. [11, 24, 20]).

Much of the conversational information exchange relies on assumptions that are not necessarily made explicit in the course of the interaction. Non-verbal signals provide an effective means to contribute to the mutual understanding of the conversation, and to update one’s knowledge without interrupting verbal presentation. In our sample dialogues, the speaker’s gesturing in Figures 1 and 2 catches the partners’ attention and also illustrates the story-line, and in Figures 3 - 5, gesturing is an effective

way to emphasize and structure the presentation. Moreover, in Figures 4 and 5, the partner's body movements also function as non-verbal feedback to the partner, and the verbal clarification in Figure 5 is nicely aligned with the original speaker's explanation. It is through this kind of verbal and non-verbal communication that the speakers construct mutual knowledge and create social bonds. This is an indication of Visual Interaction Management, i.e. controlling and coordinating the conversation by non-verbal means, without explicit verbal utterances (cf. [25]). Since communication takes place dynamically in the interactive situation, non-verbal signals provide an effective way to lead the conversation and direct the partner towards the intended interpretation without disrupting the verbal activity in communication.

The interpretation of non-verbal communicative events is related to the context in which the event occurs. The "meaning" of the gestures correlates with the communicative need on the utterance level, but it also gets interpreted in the larger context of the dialogue situation and the social activity the participants are involved in, and ultimately also includes the whole cultural context of the interlocutors. We describe the context of communicative situations in terms of activity types and the speakers' roles; cf. [29, 1]. Activity types place constraints on the speakers' behaviour when engaged in the activity and determine what are considered rational and relevant contributions in the communicative situation, i.e. appropriate with respect to the speaker's role in the conversation. The constraints further set up strong expectations on how contributions should be interpreted in a given context. Besides the activity type, the speakers' social relations also affect the interpretation of communicative behaviour: closeness of relationship between the participants creates different patterns.

# 4 Echo State Networks

Echo state networks (ESN) is a relatively novel architecture for recurrent neural networks (RNN) developed by [19]. They are capable of pattern recognition and production, and are used in the course of our work. Among the advantages of an ESN over common RNNs are the stability towards noisy inputs [37] and the efficient method to adapt the weights of the network [18]. Furthermore, ESNs are applicable in many different tasks such as classification, pattern generation, controlling tasks or language modelling [37, 38, 19, 18, 39].

As seen in Figure 6, the input layer with its $K$ neurons is fully connected to the dynamic reservoir with $M$ neurons, and the reservoir is fully connected to the output layer with $L$ neurons. The most important part of the network is the so-called reservoir. It is a collection of neurons (typically, from around ten to a few thousand in number), that are loosely connected to each other. Typically, the probability of a connection $w_{ij}$ between neuron $a_i$ and neuron $a_j$ to be set (i.e. $w_{ij} \neq 0$) in the connection matrix $W$ is around 2-10% and usually decreases with a rising number of neurons within the reservoir, whereas the connections between the in- and output

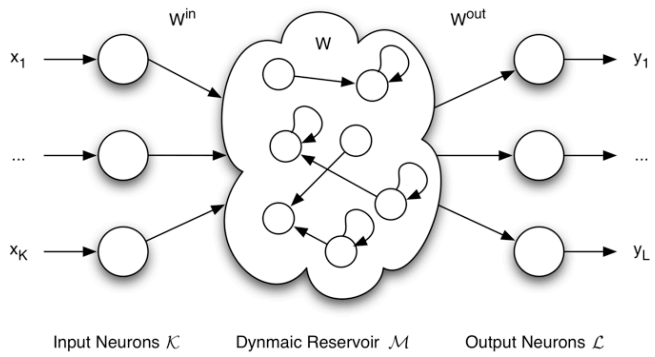


Figure 6
Schematics of an echo state network

layer with the reservoir are all set. This loose connectivity in turn leads to several small cliques of neurons that are recursively connected to each other, sensitive to a certain dynamic within the data received through the input and from other connected neurons. One of the cliques observed independently may follow a seemingly random pattern, however, if observed together with all the competing and supporting cliques within the large reservoir the reactions towards the input are anything but random. Since there are feedback and recursive connections within the reservoir, not only the input is taken into account for the output, but also the current state of each of the neurons and the history of all the inputs. Therefore, ESNs are an ideal candidate for encoding dynamic processes such as movement patterns or non-verbal utterances [37, 38, 28].

The ESN as seen in Figure 6 is simply initialized using the following parameters: since the output of the feature extraction and face tracker has eight dimensions, the same number of $K = 8$ input neurons are used [38]. Furthermore, they are connected to all the $M = 100$ neurons, with tanh as transfer function, in the reservoir with randomly initialized weights. Within the reservoir 10% of the connections are set and the weight matrix $W$ is normalized with a spectral radius of $\alpha = 0.3$ [18]. The parameter is set to a value smaller than 1 in order to achieve the targeted characteristic of behaviour in the reservoir. In order to train an ESN it is only necessary to adapt the output weights $W^{out}$ using the direct pseudo inverse method computing the optimal values for the weights from the dynamic reservoir to the output layer by solving the linear equation system $W^{out} = (S^{+}T)^{t}$, which will be further explained below. However, in general the method minimizes the distance between the predicted output of the ESN and the target signal $T$.

A network with $K$ inputs, $M$ internal neurons, and $L$ output neurons as shown in Figure 6 is considered in the following explanation. Activations of input neurons at time step $n$ are $U(n) = (u_1(n), ..., u_K(n))$, of internal units are $X(n) = (x_1(n), ..., x_M(n))$, and of output neurons are $Y(n) = (y_1(n), ...., y_L(n))$. Weights

for the input connection in an $N \times K$ matrix are $W^{in} = w_{ij}^{in}$, for the internal connection in an $M \times M$ matrix are $W = w_{ij}$, and for the connection to the output neurons in an $L \times M$ matrix are $W^{out} = w_{ij}^{out}$. The activation of internal and output units is updated according to:

$$X(n+1) = f(W^{in}U(n+1) + WX(n)), \tag{1}$$

where $f = (f_1, ..., f_M)$ are the internal neurons output sigmoid functions. The outputs are computed according to:

$$Y(n+1) = f^{out}(W^{out}X(n+1)), \tag{2}$$

where $f^{out} = (f_1^{out}, ..., f_L^{out})$ are the output neurons output sigmoid functions. A detailed description of the offline learning procedure is given below:

- Given a sequence of inputs $(u_1, ..., u_N)$ and corresponding targets $(t_1, ..., t_N)$
- Randomly generate the matrices $W^{in}$ and $W$ and scale the weight matrix $W$ such that the maximal eigenvalue $|\lambda_{max}| \leq 1$.
- Drive the network using the training data, by computing $X(n+1) = f(W^{in}U(n+1) + WX(n))$
- Collect at each time step the state $X(n)$ as a new row into a state collecting matrix $S$, and collect similarly at each time the sigmoid-inverted teacher output $\tanh^{-1}(t_n)$ into a teacher collection matrix $T$.
- Compute the pseudo inverse $S^+$ of $S$ and put $W^{out} = (S^+T)^t$

In contrast to standard feedforward neural networks such as multi layer perceptrons, the ESN incorporates previous features and states into its current state, rendering it an ideal approach for the encoding and modelling task at hand. This so-called echo state property is usually achieved by sparsely connecting the neurons of the reservoir and by scaling the weight matrix $W$ in such a way that the maximal eigenvalue $\lambda_{max} < 1$. In particular, the scaling factor $\alpha$ defines the persistence of memory [39]. Where, larger $\alpha$ values indicate that more history is considered for the current state. It is capable of modelling typical dynamics found in speech signals, and in movement during natural conversations. As is nicely shown in [28], ESNs are also capable of storing output and behavioural patterns within their dynamic reservoir. The behaviour patterns are produced by basically reversing the process direction within the network.

The Figure 6 represents an ESN. The input layer $K$ is fully connected to the dynamic reservoir $M$ via the weights stored in the matrix $W^{in}$. Within the reservoir connections are set randomly and sparsely between the neurons of reservoir $M$. Entries $w_{ij} \neq 0$ in the weight matrix $W$ correspond to connections that are set between

neuron mi and $m_j$. The output layer $L$ is again fully connected to the reservoir via the weight matrix $W^{out}$ and is adapted using the pseudo inverse method during the training of the network.

# 5 Visual Interaction Management and Activity Analysis

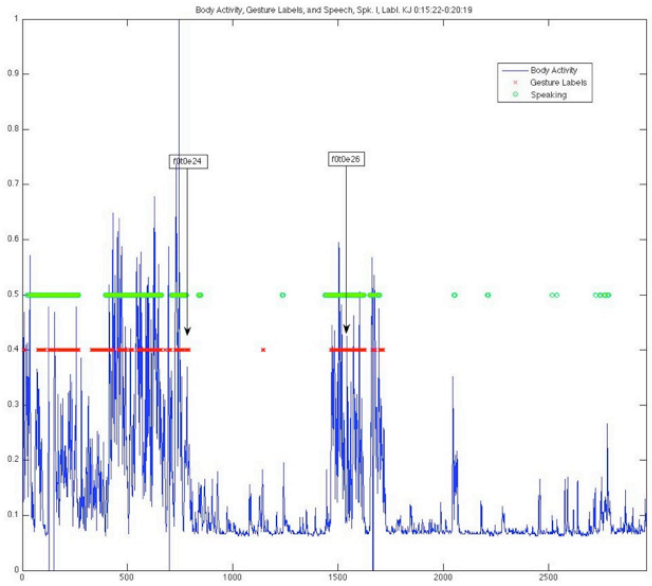


Figure 7
Body activity (blue peaks) highly correlates with speech activity (green circles the upper horizontal bar) and manual gesture labels (red crosses on the lower horizontal bar). Five minute excerpt, speaker A, annotator KJ

Conversations are full of simultaneous activity which requires subtle coordination by the participants. In this section we present signal-level analysis obtained by ESN of how the dialogue activity is seen to construct the pragmatic analysis of conversations in terms of the action to be recognized as a communicative action; and if this can be recognized on the basis of pure signal-level analysis, or if a communicative context is needed of the observed face, head, and body movement.

The faces of the participants are tracked with a standard face tracking algorithm giving the position of the faces with a very high accuracy throughout the whole conversation. The body and especially the body movement or activity is extracted automatically from a fixed area below the tracked faces. With this information and the manually labelled gestures and speech, we can analyze correlations between them.

Figure 7 displays the gesturing activity of the speaker I for the five minute segment

Table 1
Correlation coefficients for all four speakers (I, D, N, K) depict strong linear dependence between gestures (G), body activity (B), and speech (S) (values $> 0.5$)

| | **G vs. B** | **G vs. S** | **B vs. S** |
|---|---|---|---|
| **Speaker I** | 0.5836 | 0.6623 | 0.6006 |
| **Speaker D** | 0.5473 | 0.3541 | 0.209 |
| **Speaker N** | 0.3435 | 0.0595 | 0.2153 |
| **Speaker K** | 0.3276 | 0.4202 | 0.212 |

Table 2
Correlation coefficients for all four speakers (I, D, N, K) depict linear dependence between facial gestures (G), facial activity (F), and speech (S) (values > 0.3)

| | **G vs. B** | **G vs. S** | **B vs. S** |
|---|---|---|---|
| **Speaker I** | 0.3339 | 0.069 | 0.3515 |
| **Speaker D** | 0.3054 | 0.0117 | 0.3401 |
| **Speaker N** | 0.2315 | 0.1214 | 0.2553 |
| **Speaker K** | 0.3004 | 0.2966 | 0.2769 |

analysed above: bodily activity as blue peaks, speech is overlaid with green circles and gesture labels as red crosses. The identification numbers f0t0e24 and f0t0e26 identify the particular hand gesturing events and the time stamps when they occur in Figures 1 and 2: from 77.71s to 79.34s, and from 146.34s to 161.76s, respectively. The annotation analysis of the gesturing has assigned the following information to the gestures (MUMIN annotation tags): the gesturing in Figure 1 is described as SpeakerI.HandGesture, SingleHands, Centre, LeftHandDown, Single, Happy, SeqClose, TopClose, IndexOthers, while the gesturing in Figure 2 is described as SpeakerI.HandGesture, BothHands, Centre, RightHandComplex, LeftHandComplex, Repeated, TurnHold, SeqContinue, Emphasis, IndexBeats.

A clear correlation between speech, gestures, and detected activity are seen and confirmed by the correlation coefficients listed in Table 1. However, not all speakers seem to correlate as well as others (see speaker $n$ in Table 1). Speaker $n$ often moves back and forth with his body giving false activity peaks in the movement analysis as seen in Figure 8.

In a further analysis we tried to find the same strong correlates between facial gestures, speech, and facial movement or activity. However, they are not given as explicitly which can be seen in Table 2. However, correlations are still seen in the plot in Figure 9, showing the same data as Figures 7 and 8, but the movement and the gestures only correspond to the face of the speaker. This might be the result

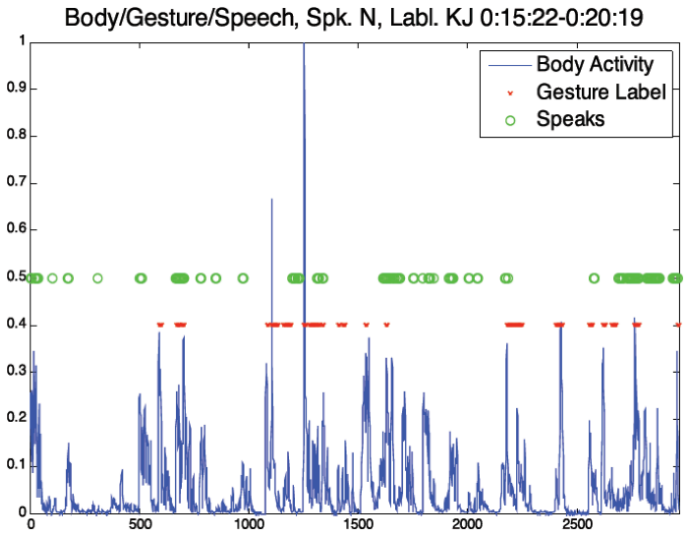


Figure 8
Body activity (blue peaks) lightly correlates with speech activity (green circles the upper horizontal bar) and manual gesture labels (red crosses on the lower horizontal bar). Five minute excerpt, speaker C, annotator KJ

of the tradeoff between detailed features and the unobtrusive and not intimidating recording setup with only one single camera. Facial activity is in this case a mixture of facial gestures such as smiling, head movement like nodding, and other biasing activities, such as drinking coffee or scratching.

Close-up views of the video examples in Figures 1 and 2 are given in Figures 10 and 11. The arrow marks the point in time of the video frame. We can see that at the still-point the body activity is not so large. This reflects the fact that the body data represent acceleration and "amount" of movement, and thus possible correlations with body positions that are actually held for some duration of time may be missing (and postures as well as a forward lean versus a backward lean). More information about the body posture would help to improve the results.

Combined figures showing the body activity and the video clip still-shot are shown in the appendix at the end of the paper.

**Conclusions and Future Work**

We have looked at the dialogue activity through bottom-up signal processing and top-down manual annotation. We showed that the two approaches can meet by visualising the activity analyses, thus providing further possibilities for experimenting with dialogue strategies and conversational control. Moreover, the results indicate close links between action and speech. We have used movement information of the face, hand, and body to estimate behavioural patterns of the conversational partici-

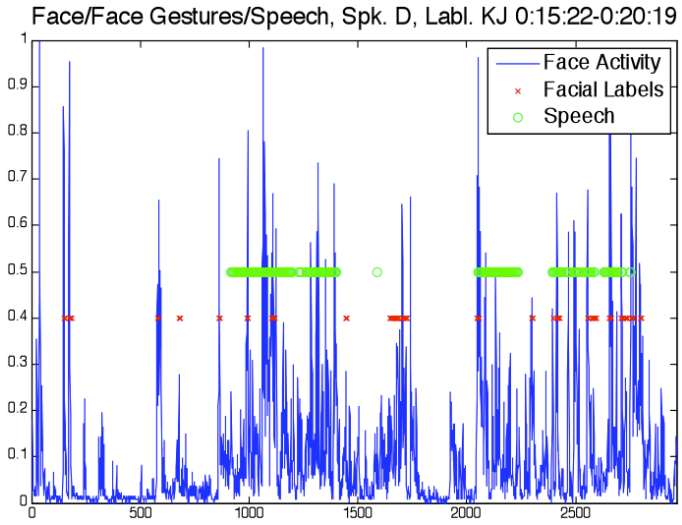


Figure 9
Face activity (blue peaks) correlates with speech activity (green circles the upper horizontal bar) and manual facial gesture labels (red crosses on the lower horizontal bar). Five minute excerpt, speaker D, annotator KJ

pants. Additionally, the results can be improved by time information e.g. about the length of particular body and hand gesture, which will be studied in the future.

The training of artificial neural networks such as Echo State networks, learning the dynamics of gestures or movement, could be used to detect different talk styles or conversational parts, like active listening, the dominant speaker, explaining something, commanding, and so on, are a matter of future work. To this end, we also tried to see if movement information of the body would provide us further understanding on turn-taking, which is an important aspect in smooth conversation management. This hypothesis is natural also from the point of view of Visual Interaction Management: the body movement is a clearly visible signal of the partner's intention, cf. [21] who noticed that in multi-party conversations head movement may signal turn-taking in a more effective manner than eye-gazing. Moreover, [8] used ESN in studying turn-taking activity, and it would be natural to see if the current data provides similar synchrony and alignment results. However, in this respect we were not successful. The reason may be that the task is simply too hard, since four people at a time are too many to model by just looking at the participants' movement behaviour. In this respect we should have more data. Also, adding more modalities to the data, like audio features, might help to estimate turn-taking behaviour; e.g. [8] used speech and were able to study turn-taking in two-people conversations.

Future work concerns further detailed analysis of social communication and synchrony between gestures and verbal utterances. We will continue the two-stage approach and integrate top-down annotations and bottom-up signal processing to learn

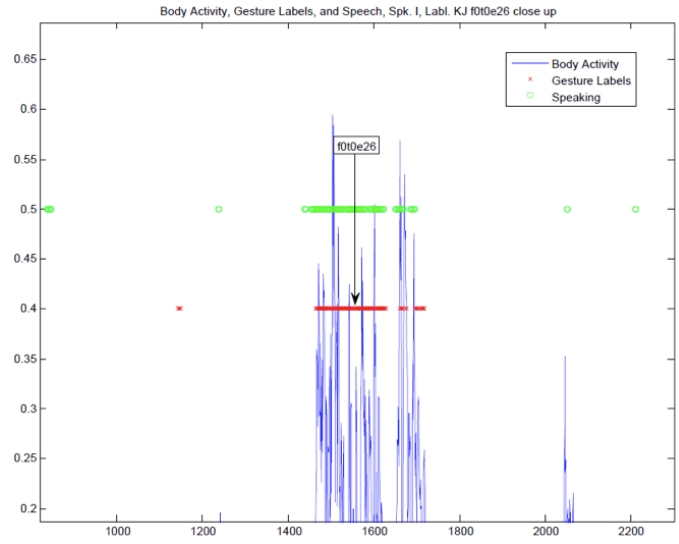


Figure 10
Close up of the speaker I's gesturing

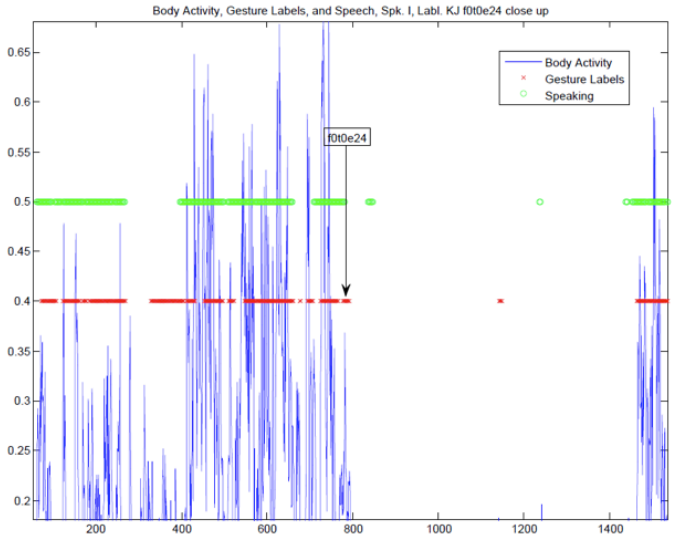


Figure 11
Close up of the speaker I's gesturing

more about how observations on signal-level match with the linguistic-pragmatic categories and human cognitive processing. We will also use more data to study

correspondences and the interplay between verbal and non-verbal interaction further. The meaning of gestures and body movement in multiparty conversational settings will be further analysed, and we also aim to develop techniques to provide basis for length and frequency studies in the timing and interplay between verbal and non-verbal signals. This will allow us to investigate Visual Interaction Management in multiparty conversations.

## References

[1] J. Allwood. Bodily communication - dimensions of expression and content. In B. Granström, D. House, and I. Karlsson, editors, *Multimodality in Language and Speech Systems*, pages 7–26. Kluwer, 2002.

[2] J. Allwood, L. Cerrato, K. Jokinen, C. Navarretta, and P. Paggio. The mumin coding scheme for the annotation of feedback, turn management and sequencing phenomena. *Multimodal corpora for modelling human multimodal behaviour. Special issue of the International Journal of Language Resources and Evaluation*, 41(3-4):273–287, 2007.

[3] E. André, T. Rist, and J. Müller. Employing ai methods to control the behavior of animated interface agents. *Applied Artificial Intelligence*, 13:415 – 448, 1999.

[4] M. Argyle and M. Cook. *Gaze and mutual gaze*. Cambridge University Press, 1976.

[5] J. Bavelas and J. Gerwing. Linguistic influences on gesture's form. *Gesture*, 4(2):157–195, 2004.

[6] M. Bennewitz, F. Faber, D. Joho, and S. Behnke. Fritz - a humanoid communication robot. In *Proceedings of the 16th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, 2007.

[7] N. Campbell and K Jokinen. Non-verbal information sources for constructive dialogue management. In *In Proc. of Language Resources and Evaluation Conference (LREC)*. ELRA, 2008.

[8] N. Campbell and S. Scherer. Comparing measures of synchrony and alignment in dialogue speech timing with respect to turn-taking activity. In *Proceedings of Interspeech*. ISCA, 2010.

[9] J. Cassell, Y. Nakano, T. W. Bickmore, C. Sidner, and C. Rich. Non-verbal cues for discourse structure. In *Proceedings of the 41st Annual Meeting of the ACL*, pages 106–115, 2001.

[10] J. Cassell, H. H. Vilhjálmsson, and T. W. Bickmore. Beat: the behavior expression animation toolkit. In *Proceedings of SIGGRAPH*, pages 477–486. ACM Press, 2001.

[11] H. Clark and E. F. Schaefer. Contributing to discourse. *Cognitive Science*, 13:259–294, 1989.

[12] M. C. Corballis. *From Hand to Mouth. The origins of language*. Princeton University Press, 2002.

[13] C. E. Douglas, N. Campbell, R. Cowie, and P. Roach. Emotional speech: Towards a new generation of databases. *Speech Communication*, 40:33–60, 2003.

[14] R. S. Feldman and B. Rim. *Fundamentals of Nonverbal Behavior*. Cambridge University Press, 1991.

[15] M. Gullberg. A helping hand? gestures, l2 learners, and grammar. In S. McCafferty and G. Stam, editors, *Gesture, Second Language acquisition and Classroom research. Applied Linguistics Series*, pages 185–210. Routledge, 2008.

[16] D. Heylen, M. Ghijsen, A. Nijholt, and R. op den Akker. Facial signs of affect during tutoring sessions. In J. Tao, T. Tan, and R. Picard, editors, *Proceedings of Affective Computing and Intelligent Interaction (ACII 2005)*, pages 24–31. Springer, 2005.

[17] E. Horvitz, C. Kadie, T. Paek, and D. Hovel. Models of attention in computing and communication: from principles to applications. *Communications of ACM*, 43(3):52–59, 2003.

[18] H. Jaeger. Tutorial on training recurrent neural networks, covering BPPT, RTRL, EKF and the echo state network approach. Technical Report 159, Fraunhofer-Gesellschaft, St. Augustin Germany, 2002.

[19] H. Jaeger and H. Haas. Harnessing nonlinearity: Predicting chaotic systems and saving energy in wireless communication. *Science*, 304:78–80, 2004.

[20] K. Jokinen. *Constructive Dialogue Management - Speech Interaction and Rational Agents*. Wiley, 2009.

[21] K. Jokinen. Gestures and synchronous communication management. In A. Esposito, N. Campbell, C. Vogel, A. Hussain, and A. Nijholt, editors, *Development of Multimodal Interfaces: Active Listening and Synchrony*, volume 5967, pages 33–49. Springer, 2010.

[22] K. Jokinen, P. Paggio, and C. Navarretta. Distinguishing the communicative functions of gestures. an experiment with annotated gesture data. machine-learning for multimodal interaction. In *Proceedings of 5th International Workshop, MLMI 2008*, pages 38–49. Springer, 2008.

[23] K. Jokinen and M. Vanhasalo. Stand-up gestures – annotation for communication management. In *Proceedings of the Multimodal Workshop at Nodalida Conference*, 2009.

[24] Y. Katagiri. Interactional alignment in collaborative problem solving dialogues. In *Proceedings of 9th International Pragmatics Conference*, 2005.

[25] A. Kendon. *Gesture: Visible action as utterance*. Cambridge, 2004.

[26] M. Kipp. *Gesture Generation by Imitation: From Human Behavior to Computer Character Animation*. Dissertation.com, Boca Raton, Florida, 2005.

[27] H. Koiso, Y. Horiuchi, S. Tutiya, A. Ichikawa, and Y. Den. An analysis of turn taking and backchannels based on prosodic and syntactic features in japanese map task dialogs. *Language and Speech*, 41(3-4):295–321, 1998.

[28] A. F. Krause, B. Blaesing, V. Duerr, and T. Schack. Direct control of an active tactile sensor using echo state networks. In H. Ritter, G. Sagerer, R. Dillmann, and M. Buss, editors, *Proceedings of 3rd International Workshop on Human-Centered Robotic Systems (HCRS'09)*, Cognitive Systems Monographs, pages 11–21. Springer, 2009.

[29] S. Levinson. Activity types and language. In Drew P. and Heritage J., editors, *Talk at work*, pages 67–107. Cambridge University Press., 1992.

[30] J. C. Martin, R. Niewiadomski, L. Devillers, S. Buisine, and C. Pelachaud. Multimodal complex emotions: gesture expressivity and blended facial expressions. *International Journal of Humanoid Robotics. special issue on "Achieving Human-Like Qualities in Interactive Virtual and Physical Humanoids"*, 3(3):269–291, 2006.

[31] Y. Nakano and T. Nishida. Attentional behaviours as nonverbal communicative signals in situated interactions with conversational agents. In T. Nishida, editor, *Conversational Informatics*, chapter 5, pages 85–102. Wiley, 2007.

[32] E. Padilha and J. Carletta. Nonverbal behaviours improving a simulation of small group discussion. In *Proceedings of the 1st Nordic Symposium on Multimodal Communication*, pages 93–105, 2003.

[33] C. Pelachaud, V. Carofiglio, B. D. Carolis, F. de Rosis, and I. Poggi. Embodied contextual agent in information delivering application. In *Proceedings of the First international joint conference on Autonomous agents and multiagent systems (AAMAS'02)*, pages 758–765. ACM Press, 2002.

[34] M. Pickering and S. Garrod. Towards a mechanistic psychology of dialogue. *Behavioral and Brain Sciences*, 27:169–226, 2004.

[35] M. Rehm, Y. Nakano, E. André, T. Nishida, N. Bee, B. Endrass, M. Wissner, A. Lipi, and H.-H. Huang. From observation to simulation: generating culture-specific behavior for interactive systems. *AI Soc.*, 24(3):267–280, 2009.

[36] T. Rist, E. André, and S. Baldes. A flexible platform for building applications with life-like characters. In *Proceedings of the 8th international conference on Intelligent user interfaces (IUI '03)*, pages 158–168. ACM Press, 2003.

[37] S. Scherer, M. Oubbati, F. Schwenker, and G. Palm. Real-time emotion recognition from speech using echo state networks. In *Proceedings of the 3rd IAPR workshop on Artificial Neural Networks in Pattern Recognition (ANNPR'08)*, pages 205–216, Berlin, Heidelberg, 2008. Springer.

[38] S. Scherer, F. Schwenker, W. N. Campbell, and G. Palm. Multimodal laughter detection in natural discourses. In H. Ritter, G. Sagerer, R. Dillmann, and M. Buss, editors, *Proceedings of 3rd International Workshop on Human-Centered Robotic Systems (HCRS'09)*, Cognitive Systems Monographs, pages 111–121. Springer, 2009.

[39] M. H. Tong, A. D. Bickett, E. M. Christiansen, and G. W. Cottrell. Learning grammatical structure with echo state networks. *Neural Networks*, 20:424–432, 2007.

[40] N. Ward and W. Tsukahara. Prosodic features which cue back-channel responses in english and japanese. *Journal of Pragmatics*, 23:1177–1207, 2000.

## Appendix

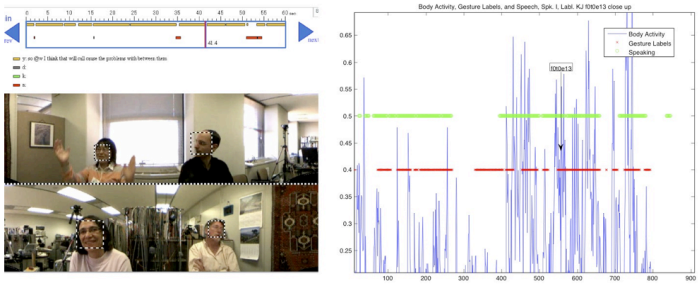


Figure 12
Speaker I

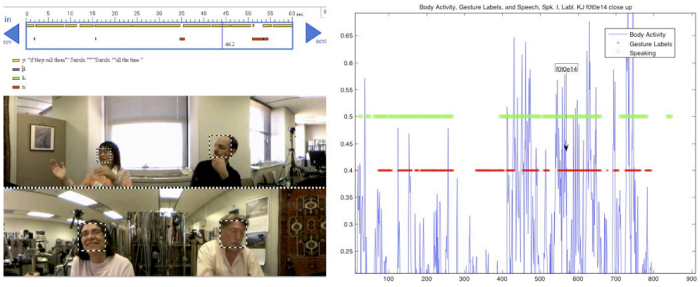


Figure 13
Speaker I

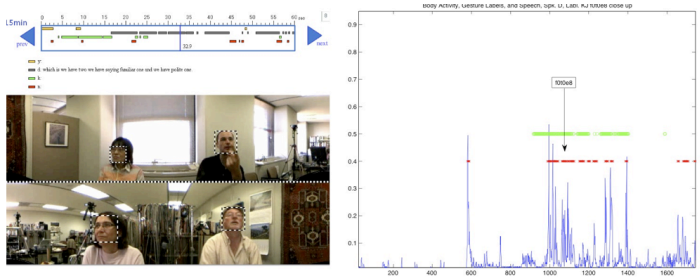


Figure 14
Speaker D

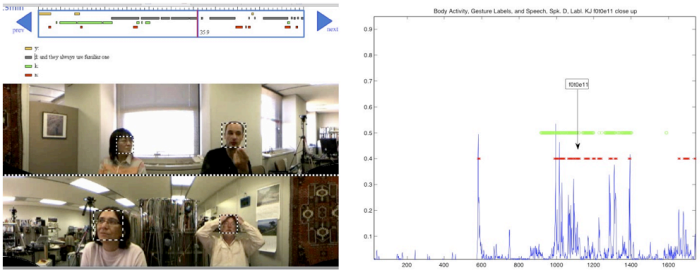


Figure 15
Speaker D

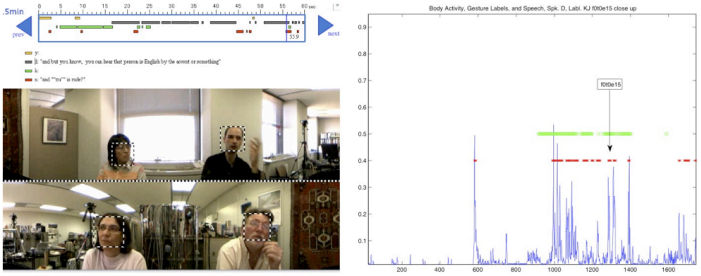


Figure 16
Speaker D

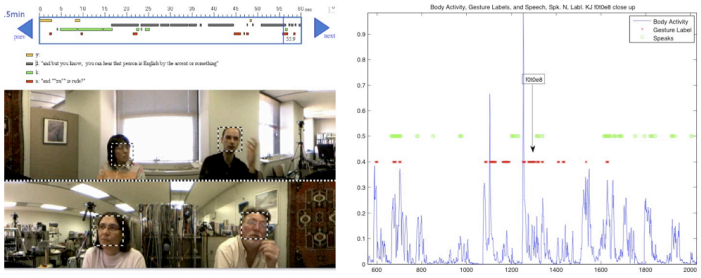


Figure 17
Speaker N