Generating Listening Behaviour

Dirk Heylen, Elisabetta Bevacqua, Catherine Pelachaud, Isabella Poggi, Jonathan Gratch, and Marc Schröder

Abstract In face-to-face conversations listeners provide feedback and comments at the same time as speakers are uttering their words and sentence. This 'talk' in the backchannel provides speakers with information about reception and acceptance – or lack thereof – of their speech. Listeners, through short verbalisations and non-verbal signals, show how they are engaged in the dialogue. The lack of incremental, real-time processing has hampered the creation of conversational agents that can respond to the human interlocutor in real time as the speech is being produced. The need for such feedback in conversational agents is, however, undeniable for reasons of naturalism or believability, to increase the efficiency of communication and to show engagement and building of rapport. In this chapter, the joint activity of speakers and listeners that constitutes a conversation is more closely examined and the work that is devoted to the construction of agents that are able to show that they are listening is reviewed. Two issues are dealt with in more detail. The first is the search for appropriate responses for an agent to display. The second is the study of how listening responses may increase rapport between agents and their human partners in conversation.

1 Introduction

In many books and papers, the process of communication is schematically depicted with a speaker who is active in the speech process and the listener who is involved in passively perceiving and understanding the speech. According to Bakhtin (1999) linguistic notions such as 'the "listener" and "understander" (partners of the "speaker") are *fictions* which produce a 'distorted idea' of the process of speech communication.

AQ1

D. Heylen (✉)
University of Twente, Enschede Faculty of Electrical Engineering Mathematics and Computer Science, The Netherlands
e-mail: heylen@ewi.utwente.nl

P. Petta et al. (eds.), *Emotion-Oriented Systems*, Cognitive Technologies,
DOI 10.1007/978-3-642-15184-2_17,

> One cannot say that these diagrams are false or that they do not correspond to certain aspects of reality. But when they are put forth as the actual whole of speech communication, they become a scientific fiction. The fact is that when the listener perceives and understands the meaning (the language meaning) of speech, he simultaneously takes an active, responsive attitude toward it. He either agrees or disagrees with it (completely or partially), augments it, applies it, prepares for its execution, and so on. And the listener adopts his responsive attitude for the entire duration of the process of listening and understanding, from the very beginning – sometimes literally from the speaker's first word. [. . .] Any understanding is imbued with responsive and necessarily elicits it in one form or another: the listener becomes a speaker.

Moreover, Bakhtin claims, any speaker is in a sense also a respondent. It seems then that when one attempts to create virtual humans that act as listeners, one is engaged in writing science fiction in the second degree unless one takes the dialectic between speaking and listening by listeners and speakers, respectively, into account.

In order to create agents that can listen to the speech of the humans they interact with, we need to have a proper understanding of what constitutes listening behaviour and how communication in general proceeds. In the first section of this chapter we will introduce the major terms and concepts that are relevant for understanding what listeners do. After this we can turn to the many challenges that are involved in creating conversational agents that have similar abilities. We will focus on two issues that have been considered in the virtual agent literature. The first involves the use of conversational agents or synthesised vocal expressions in the search for listener signals. The second point concerns the use of 'active' listening behaviours to create rapport with the human interlocutor.

2 Understanding Communication

Bakhtin is not the only one who makes the point that listeners are not just passive recipients of messages emitted by a speaker. Conversation has been characterised as a collaborative activity, an interactional achievement or a joint activity by researchers such as Gumperz (1982), Schegloff (1982) and Clark (1996). By using the term interactional achievements Schegloff highlights the fact that conversations are incrementally accomplished and they involve dependency of the actions of one participant on the actions of the other and vice versa. The term joint activity is used by Clark to emphasise that it is only when the participatory actions of the different participants are seen *together* that one can talk about a conversation.

Communicative actions of one participant implicate the others in many ways. A typical communicative action is normally produced with the intention that one or more other participants (the addressees, the audience, the 'listeners') attend to them, are able to perceive them, recognise the behaviour as an instance of a communicative action, try to understand them and possibly act upon them in one way or another, preferably with the effect that the producer of the communicative action had intended to achieve. If these conditions are not met the action will fail to be 'happy' in Austin's term (Austin, 1962) or will not be 'felicitous' (Searle, 1969).

The success of a communicative action thus depends on the states of mind and the behaviours of the other participants during the preparation and execution and ending of the communicative behaviours. As Schegloff and others have pointed out, the behaviours of the other participants not only determine success but they may also influence and change the execution of the communicative actions *as they are being produced*, because the producer of the action will take notice of how the audience receives and processes the actions and also of the other reactions they invoke. A nice example is provided by Goodwin (1984) who defines as a principal rule in face-to-face conversation that 'When a speaker gazes at a recipient that recipient should be gazing at him. When speakers gaze at nongazing recipients, and thus locate violations of the rule, they frequently produce phrasal breaks, such as restarts and pauses, in their talk' (Goodwin, 1984, p. 230). Similarly, Kraut et al. (1982) conducted some experiments which made it clear how speakers adjust the informational density of their talk depending on the kind and amount of verbal feedback they receive from listeners. Speakers may also monitor listeners for the various actions besides listening that they are involved in. An experiment set up by Clark and Krych (2004), for instance, made it clear that in a collaborative task, not being able to monitor the other's face and eye gaze had less of an effect than not being able to see the other's workspace and what activity was being performed. Clearly, the setting and task involved in the conversation may assign different priorities to what kind of feedback of the interlocutors is important to monitor and what effect this has on the way the conversation proceeds.

We can picture the interaction between actions of the participants in conversation in a first, simple diagram (Fig. 1) which is only slightly more complicated than the fictions Bakhtin was referring to but it tries to show something more of the dialogical nature of conversation.

For the sake of simplicity, assume that a conversation takes place between two persons (*x* and *y*). Given that some conversational action (CA1) is performed by one of them (say *x*), as indicated by the top left corner (A) of this diagram, the other person (*y*) is supposed to perceive and interpret this action, as indicated by the top right corner (B). We will summarise the various actions that this involves using the term 'perceive', which is taken from the classical notion in artificial intelligence

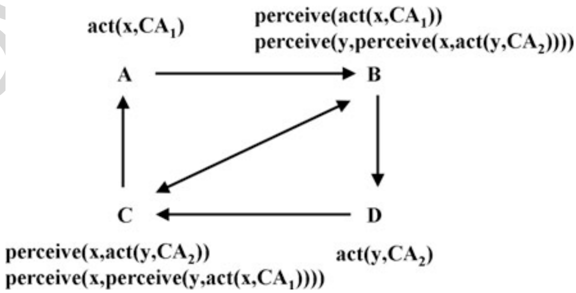

Fig. 1 Picturing conversation as an interactional achievement

that an intelligent agent is involved in perception–decision–action loops. This may prompt this person (y) (i.e. lead y to decide) to produce certain actions (CA2 in the bottom right corner, D). These actions in turn can communicate something to the producer of CA1 (x) about the reception and up-take of the production of CA1 by y (bottom left corner, C) which may either change the execution of action CA1 or prompt a new action. The behaviours that make up the act of perception of CA1 by y (B) may themselves be observable to x who is monitoring them, hence the arrow connecting corner B with C. Vice versa, the actions that go into the perception of CA2 by x may also be observable to y. Actions by one thus elicit actions by the other in reply.

So far, only general terms such as 'communicative action', 'producer' and 'recipient' and 'perceiver' were used because any action could enter these perception–action loops. Therefore, also the time scale was left unspecified. The diagram can be instantiated in many different ways. For instance, the communicative action CA1 by x could be the utterance of statement, which makes x a *speaker* during which y, the *listener*, attending to the speech, shows a puzzled face (CA2) accompanied by a vocalisation 'oh' with a rising intonation. This verbal and non-verbal *feedback* in the *backchannel*, which is monitored by the speaker x, may prompt x to enter into reformulation mode or to speak up. All of this can happen almost instantaneously, slowed down only by the limits of the speed of light, sound and neurons firing but also sped up through the force of anticipation by both x and y which makes it even possible for the agents to run ahead of events. At any given time, there will be multiple instantiations of the schema active as participants can communicate with different modalities in parallel or because one can view the process as operating on different levels as will be pointed out below.

Another common instantiation is the case where someone (x) produces a speech act (CA1), which is attended to and interpreted by y who decides to offer a speech act (CA2) in reply, after which x responds by producing a new speech act (CA1′). The two participants take alternating turns and each next utterance is a reply to the previous one forming adjacency pairs as they are commonly called in the tradition (Schegloff and Sacks, 1973) of conversation analysis.[1]

A third common instantiation has been labelled *interactional synchrony*. It was first described by Condon and Ogston (1966) and an episode in a conversation was analysed in detail by Kendon (1970). The term refers to the case where the flow of movements of the listener are rhythmically coordinated with those of the speaker. Other forms of coordination have been called mimicry (Chartrand and Bargh, 1999) and mirroring (LaFrance, 1979; Lafrance and Ickes, 1981). Hadar and colleagues (1985) report that approximately a quarter of all the head movements of the listeners in the conversations they looked at occurred in sync with the speech of the interlocutor. Interestingly, McClave (2000) notes that (many of) these kinds of movements may be elicited by the speaker.

[1] Goffman (1976) provides a very insightful analysis of this process of replies and responses.

> Many instances of backchanneling were assumed to be internally motivated; i.e. the listener backchanneled when he or she felt like it. Microanalysis of speaker head movements in relation to listener head movements reveals that what were heretofore presumed to be spontaneous, internally motivated, listener responses are actually responses to the speaker's nonverbal requests for feedback. These requests are in the form of up-and-down nods, and listeners recognize and respond to such requests in a fraction of a second.

Again, this shows the dependence of an action by one participant on the action of another, the back-and-forth of eliciting actions and responses.

Clearly, what has been understood above by a communicative action is very broad. It may involve consciously produced linguistic actions but also actions that were not meant to be communicative by the producer but that still provide information to the recipient. The communicative behaviours may 'signal' in various ways: symbolically, indexically, iconically or through inference.

In the following paragraphs we present a variety of instantiations of this schema as we discuss some central theoretical notions and some common ways in which the interactions between participants in conversation proceed. We will detail how actions of one participant call forth or intend to call forth actions of others and what kinds of responses one can distinguish.

2.1 Speech Acts

The crucial insight that speech act theory (Austin, 1962; Searle, 1969) has emphasised is that 'language is used for getting things done'. Typically, in the case of language, these things implicate the person or persons to which the utterance is being addressed. From a speech act perspective, any utterance is some kind of invitation to the addressees to participate in a particular configuration of actions: Attend to what is being said, try to figure out what is meant and carry out what was intended by the speaker, which could range from updating a belief state, to feeling offended, or closing the window. Speech act theory focusses on the perspective of the speakers and their *intentions* which implicate the audience in that an utterance is primarily intended to get the audience to recognise the speaker's meaning: 'To say that a speaker meant something by X is to say that the speaker intended the utterance of X to produce some effect in the audience by means of the recognition of this intention.' This is essentially Grice's definition (Grice, 1975b). Another way in which the perspective of the speaker comes to the fore is in the way that Grice (1975a) formulates his maxims of cooperative behaviour (be relevant, be conspicuous, etc.) in terms of what the speaker should and should not do. All of these maxims indirectly take listeners into account as they urge the speaker to keep them in mind for the sake of cooperation.

As with any event, a speech event can be described in several ways. One might say that in describing a particular situation the speaker was 'stuttering', 'trying to say something in English', 'trying to propose', 'making a fool of himself', etc. By using the word 'stuttering' one is referring to an aspect of the production and vocalisation process. The second characterisation points out that the vocalisations were not random but attempt to construct an English sentence. The third describes the intention behind the action and the last the effect it may have achieved on the other participants, the observers or those that have heard about the event.

Austin (1962) proposed some different terms to distinguish the levels in the speech event. The uttering itself he called the locutionary act. The act of getting the audience to recognise what is intended is called the illocutionary act (the speaker tries to make it clear that the utterance is intended as a promise, for example). The effects the execution of the speech act has on the audience are called the perlocutionary effects. The acts that caused these effects were the perlocutionary acts. Note that not all of the effects may have been intended. For instance, if the speaker is not aware that the action promised is not something the audience wants, then the promise may actually turn out to be a threat.

In Clark's framework (Clark, 1996), a speaker acts on four levels. (1) A speaker executes a behaviour for the addressee to attend to. This could be uttering a sentence but also holding up your empty glass in a bar (to signal to the waiter you want a refill). (2) The behaviour is presented as a signal that the addressee should identify as such. It should be clear to the waiter that you are holding up the glass to signal to him and not just because of some other reason. (3) The speaker signals something which the addressee should recognise. (4) The speaker proposes a project for the addressee to consider (believe what is being said, except the offer, execute the command, for instance). In this formulation of levels, every action by the speaker is matched by an action that the addressee is supposed to execute: Attend to the behaviour, identify it as a signal, interpret it correctly and consider the request that is made. If one considers the diagram above, one could say that instead of one arrow going from A to B there are four. Also, the arrow should be considered both from the perspective of the speaker and the recipient.

2.2 Monitoring and Feedback

If we take the perspective of the listener, we can make a similar distinction in four levels on which the listener can provide feedback. Allwood (1993), for example, put forward a distinction of the following four basic communicative functions on which the interlocutor can give feedback:

1. Contact (i.e. whether the interlocutor is willing and able to continue the interaction)
2. Perception (i.e. whether the interlocutor is willing and able to perceive the message)
3. Understanding (i.e. whether the interlocutor is willing and able to understand the message)
4. Attitudinal reactions (i.e. whether the interlocutor is willing and able to react and (adequately) respond to the message, specifically whether he/she accepts or rejects it).

Important for all the parties in the cooperative undertaking that is conversation is to know that common ground has been established, that the addressee understands what the speaker intended with the talk produced and the speaker knows that the intentions were achieved. So the feedback that is voluntarily or involuntarily provided by listeners is monitored by the speakers in order to get closure on their actions, i.e. in order to know to what degree the intended actions were successful. Goodwin's rule – whenever a speaker looks at his audience, the audience should look at the speaker – provides a basic example of this need to check for contact and perception. By monitoring the behaviour of the other participants, a speaker can thus derive information about such elements as attention, perception, understanding and the willingness to engage and accept or reject collaboration. Some of the information derives from the actions of listeners that go into perception of the signals (such as their gaze telling something about the focus of attention) but other behaviors may be explicit signals of understanding and agreement or lack thereof through facial expressions or small non-disruptive interjections. This we will discuss in Sect. 2.3. Also the way the utterances are taken up by subsequent actions are informative and provide the speaker with feedback on the conversational moves, of course.

Several conversational actions are conventionally dedicated to establish 'grounding' (the mutual belief by the partners in conversation that they have understood what the contributor meant; Clark and Schaefer (1991)). In Clark and Schaefer, a discourse model is presented in which it is assumed that the presentation phase of the speaker is paralleled with an acceptance phase by the recipient which is essential for grounding. Either following, in the next moves or by behaviours during the production of communicative actions by the speaker. Obvious signs of neglect of attention or signs of difficulty in understanding will yield reparative actions by the speaker. Positive signs indicating attention, perception, understanding, processing (understanding, agreement, willingness, etc.) will lead the speaker to assume the message has been grounded or successfully executed on all the relevant levels.

> The acceptance phase is usually initiated by B giving A evidence that he believes he understands what A mean by *u*. B's evidence can be of several types. He can say that he understands, as with *I see* or *uh huh*. Or he can *demonstrate* that he understands, as with a paraphrase, or what it is he heard, as with as with a verbatim repetition. Another is by showing his willingness to go on. The least obvious way is by showing continued attention. (Clark and Schaefer, 1991)

The acceptance phase itself consists of the presentation of a contribution to which the original presenter can react with an accepting contribution, illustrating another way to describe some of the loops presented in Fig. 1.

One type of accepting contribution Clark and Schaefer call *acknowledgements*, which are 'expressions such as *mhm*, *yes*, and *quite* that are spoken in the background, or gestures such as head nods and smiles'. These are commonly called *backchannels*.

2.3 Backchannels

Yngve (1970) is generally credited for having introduced the term. His characterisation is this. Note how it repeats some of the points made by Bakhtin.

> One should hasten to point out that the distinction between having the turn or not is not the same as the traditional distinction between speaker and listener, for it is possible to speak out of turn, and it is even reasonably frequent that a conversationalist speaks out of turn. In fact, both the person who has the turn and his partner are simultaneously engaged in both speaking and listening. This is because of the existence of what I call the back channel, over which the person who has the turn receives short messages such as 'yes' and 'uh-huh' without relinquishing the turn. The partner, of course, is not only listening, but speaking occasionally as he sends the short messages in the back channel. The back channel appears to be very important in providing the monitoring of the quality of communication.

Several authors, Duncan and Fiske for instance (Duncan and Fiske, 1977), have used the term *backchannel* but the interpretation of the term shows some variation. In part, the instability of the meaning can be traced back to the difficulty in specifying the denotation of some terms that one commonly encounters in the definition of *backchannel*, such as *turn* (or *floor*), *listener* (or *hearer*, *auditor*, *recipient*) and *speaker*. Another difficulty in defining the term is that there is quite some variation in the kinds of behaviours and in the kinds of functions that 'listeners' produce as 'feedback.' The term *backchannel* is sometimes reserved for a particular subset of these behaviours and sometimes taken to include a much wider range of behaviours.

Some authors use other terms to refer to similar phenomena sometimes restricting the scope to a particular class of listener responses. Kendon (1967) introduced the term accompaniment signals for 'short utterances that the listener produces as an accompaniment to a speaker, when the speaker is speaking at length' which he divides into two groups: attention signals (in which one appears to signal no more than that one is attending) and assenting signals that express 'point granted' or 'agreement'. Rosenfeld (1987) uses the general term *listener response*. A related concept is that of *acknowledgement token* as used by Jefferson (1984) or *continuers* from Schegloff (1982). Schegloff reflects on the use of 'uh-huh' as a signal of attention, which makes sense only if attention is somewhat problematic. Therefore this attention-signalling function of an 'uh-huh' or a head nod becomes apparent only if it is in response to an extended gaze by the speaker or a rising intonation soliciting some sign of attention, interest or understanding (Schegloff, 1982, p. 79). In other cases, the term continuer may be appropriate, according to Schegloff.

> Perhaps the most common usage of 'uh huh', etc. (in other environments than after yes/no questions) is to exhibit on the part of its producer an understanding that an extended unit of talk is underway by another, and that it is not yet, or may not yet be), complete.

AQ4

The responses that listeners provide to speakers falling under the general cover-all term backchannel (as used by Duncan and Niederehe, 1974) can thus have many functions, depending on the context. In the following section we will look at how some function/form relations can be identified by having people rate different samples, amongst others created by synthesis, using an embodied conversational agent.

2.3.1 Turn-Taking

In the discussion of the schema presented above, an interpretation of the schema was pointed out where a communicative action by one agent was followed by a communicative act by the other agent in the next turn. An important decision that a conversational agent needs to make is when to start speaking and when to stop and listen. So how do participants in a conversation decide when to speak and when to keep quite? Sacks et al. (1974) propose a simple systematic that says that in general a speaker can select the next speaker (for instance by asking a question to a particular person), or that the next speaker can self-select. This view on turn-taking has been criticised by various researchers. One point that is often made is that it is not very contentful. From a general characterisation of turn-taking that should apply to any conversational setting this is probably what is to be expected. The question can also be answered in another way. Instead of taking a structuralist point of view, one can also take the stance of the individual agent. In the same general mode (but now using intentional terms which conversational analysts avoid to invoke) the following could be said to hold: An agent decides to speak when the reasons for speaking outweigh the reasons for not speaking and vice versa, an agent decides not to speak when the reasons for not speaking outweigh the reasons for speaking. Now the question is what are the reasons that play a role in this decision-making process. One can imagine that the factors that play a role are enormously varied and depend a lot on the precise circumstances. Some reasons for speaking that you may have encountered personally are as follows:

1. You have something you would very much like to say.
2. You have just been asked a question and feel the pressure to answer.
3. The current speaker is about to say something embarrassing and you decide to interrupt to save the speaker from loss of face.
4. The current speaker is looking for a word and you help out, by suggesting the word you think the speaker is looking for.
5. You need something done by someone else and talking seems the best way to accomplish this.
6. There is an awkward silence and you ask your guests whether they have already planned where to go on vacation.

Some reasons you may have experienced for not claiming the turn are as follows:

1. You have nothing to say.
2. You are too embarrassed to speak.
3. Someone else is speaking and you need to hear what is being said.
4. You are afraid to say something that will hurt someone's feelings.
5. You would like to say something but the chairperson in the meeting first gives the turn to another participant.
6. You are a suspect in a police investigation; anything you say might be used against you.

7. You provide an accompaniment signal and wait for the current speaker to reach the end of a phonemic clause, i.e. the end of an informational unit, where you think it is no longer impolite to interrupt.

This huge diversity of reasons can be classified into different groups. Some have to do with the business or the task that is being carried out through conversation (task goals); others concern the feelings of the participants, the social conventions (ritual constraints in Goffman's terms (1976) and others seem to operate to make conversations work (system constraints, again using Goffman's terminology). In the following sections, we will not dwell on these issues in detail, but clearly, when designing conversational agents that show the appropriate listening behaviours, one needs to take into account the way they signal they want to continue as listeners or how they display they want to take up the speaking role; Duncan and Niederehe, 1974).

2.4 In Summary

Listeners are not merely passively absorbing what a speaker is saying. They are involved in a number of activities: attending to the actions of the speaker to see what actions the speaker elicits/evokes from them in response, showing speakers that they are attending (implicit feedback) and providing explicit feedback in all kinds of forms. As Fig. 1 shows there is a constant back and forth between the various participants in a conversation where some behaviour by one participant elicits a reaction by the other which is monitored and responded to almost instantaneously. The challenges for building embodied conversational agents are thus manifold. The agent should be able to monitor and interpret the utterances of the human interlocutor 'on the fly'. AQ5 It should be able to detect that the appropriate points were a signal of attention or of agreement needed, being careful in its timing so as not to disrupt the flow of conversation. The agent should have a repository of behaviours it can execute with all kinds of shades of meaning represented in line with its goals in the conversation and its synthetic personality.

In the following sections we will sketch some work that is currently on its way to create embodied conversational agents that can give the appearance that they know how to listen. In Sect. 3 we report on work that uses embodied agents to build up a library of function/form mappings. Ultimately, the aim is to build engaging agents that people like to interact with. In Sect. 4 we report on ongoing work that measures the effects of the display of appropriate listening behaviours by agents on the sense of engagement and rapport that is experienced by the human interlocutor.

3 Artificial Stimuli and Expression Libraries

The variety of behaviours that listeners display during face-to-face dialogues is very large. The functions that they serve are also multiple. By gazing at the speaker a listener signals attention and that the communication channels are open (Kendon, 1967). By nodding the listener may acknowledge that he has understood what the speaker wanted to communicate. A raising of the eyebrows may show that the listener thinks something remarkable is being said (Ekman, 1979; Chovil, 1991) and by moving the head into a different position the listener may signal that he wants to change roles and say something himself (Duncan and Niederehe, 1974; McClave, 2000). It was already indicated that the behaviours that listeners display are relevant to several communication management functions such as contact management, grounding, up-take and turn-taking (Allwood et al., 1993; Yngve, 1970; Poggi, 2007). They are not only relevant to the mechanics of the conversation but also to the expressive values: the attitudes and affective parameters that play a role. These attitudes can be related to a whole range of aspects, including epistemic and propositional attitudes such as believe and disbelieve but also affective evaluations such as liking and disliking (Chovil, 1991).

Some authors have investigated whether these differences in functions correlate with differences in form. Rosenfeld and Hancks (1980) made a start to determine which nonverbal behaviors of listeners were signalling either attention, understanding or agreement by having independent observers rate 250 listener responses on each of the three dimensions. They found that judgements of ‘agreement’ were associated with complex verbal listener responses and multiple head nods. Contextually, this occurred when the responses followed the speaker pointing the head in the direction of the listener. Signalling understanding was associated with more subdued forms such as repeated small head nods prior to the speaker finishing a clause. Expressions were rated highest as signalling attention when the listener ‘leaned forward prior to the speaker’s juncture, audibility of verbal listener response after the juncture, and initiation of gesticulation by the speaker after the juncture but prior to resuming speech’ (Rosenfeld, 1987).

Some important characteristics of expressive communicative behaviours are that a behaviour can signal more than one function at the same time and that behaviours may serve different functions depending on the context. In order to create conversational agents that display the appropriate behaviours in the right context it is important to get more insight into the various behaviour to function mappings. Besides looking at naturally occurring contexts, to investigate the relation between form and function, one can also get more insight into what (combinations of) expressions can be used to express what kind of information by generating artificial stimuli that are judged by people. In the following sections two such studies are presented.

3.1 Facial Expressions

In studies by Bevacqua et al. (2007) and Heylen et al., (2007a) a generate and evaluate procedure was used where people were asked to label short movies of the Greta agent displaying a combination of facial expressions. The experiments were conducted to find some prototypical expressions for several feedback functions and to gain insight into the way the various components in the facial expression contribute to its functional interpretation.[2] In particular, the aim of these experiments was to get a better understanding of

- the expressive force of the various behaviours,
- the range and kinds of functions assigned,
- the range of variation in judgements between individuals and
- the nature of the compositional structure (if any) of the expressions.

A lot has been written about the interpretation of facial expressions. This body of knowledge can be used to generate the appropriate facial expressions for a conversational agent. However, there are many situations for which the literature does not provide an answer. This often happens when one needs to generate a facial expression that communicates several meanings from different types of functions: show disagreement and understanding at the same time, for instance. The literature may provide certain pointers to expressions for each of the functions separately, but the way they should be combined may not be so easy. In another way, we know that eyebrow movements occur a lot in conversations with many different functions. The question that arises in this case is whether it makes sense to distinguish them in the way they are performed and the timing of execution or the co-occurrence with other behaviours.

In the studies, the authors looked for expressions for the following functions: *agree, like, understand, disagree, dislike, disbelieve, don't understand* and *not interested*. In the first experiment, reported in Bevacqua et al. (2007), it was found that users could easily determine when a context-free signal conveys a positive or a negative meaning. A first question that was explored in the second test was whether it is possible to find a prototypical signal (or a combination of signals) for each meaning. Is there a signal more relevant than others for a specific meaning or can a single meaning be expressed through different signals or a combination of signals? The hypothesis was that for each meaning, one can find a prototypical signal which could be used later on in the implementation of conversational agents.

A second question was in what way combinations of signals alter the meaning of single backchannel signals. It was conjectured that adding a signal to another could significantly change the perceived meaning. In the study reported on in Heylen et al. (2007a), 60 French subjects were involved in the experiment. They were divided into two groups, each of which judged about half of the movies. The test used the 3D agent, Greta (Pelachaud and Bilvi, 2003). Participants were presented 21 movies. Table 1 shows the signals, chosen from those proposed by Allwood and Cerrato, (2003) and Poggi (2007), that were used to generate the movies.

The meanings the subjects could choose from were *agree, disagree, accept, refuse, interested, not interested, believe, disbelieve, understand, don't understand, like, dislike.*

[2] Similar experiments were reported on in Heylen et al., (2007b) and Heylen, (2007).

Table 1 Backchannel signals

1. Nod	8. Raise eyebrows	15. Nod and raise eyebrows
2. Smile	9. Shake and frown	16. Shake, frown and tension[a]
3. Shake	10. Tilt and frown	17. Tilt and raise eyebrows
4. Frown	11. Sad eyebrows	18. Tilt and gaze right down
5. Tension[a]	12. Frown and tension[a]	19. Eyes wide open
6. Tilt	13. Gaze right down	20. Raise left eyebrows
7. Nod and smile	14. Eyes roll up	21. Tilt and sad eyebrows

[a]The action *tension* means tension of the lips

The list of possible meanings was proposed to the participants who, after each movie and before moving on, could select the meanings that they thought fitted the backchannel signal best. Participants were told that Greta would display backchannel signals as if Greta was talking to an imaginary speaker. This context was provided to make participants aware that they were evaluating backchannel signals. The signals were shown once, randomly: a different order for each subject.

The most significant results for each of the functions were the following.

Agree. When displayed on its own, *nod* proved to be very significant since every participant answered ‘agree’. *Nod and smile* and *nod and raise eyebrows* also scored highly as backchannel signals of agreement. On its own, a *smile* does not associate with ‘agreement’, though. Similar results were obtained for the meaning of **Accept**.

Like. Two signals conveyed the meaning ‘like’: *nod and smile* and *smile*.

Understand. Thirteen out of 30 subjects associated a nod with ‘understand’, 16 of them paired *nod and smile* with this meaning and 17 found that *nod and raise eyebrows* could mean ‘understand’. *Raise eyebrows* on its own is not associated with understanding as only one subject judged it as such.

Disagree. The signal *shake* is labelled by every subject as meaning ‘disagree’. The combination of *shake and frown and tension* is also highly recognised as ‘disagree’. Also the combination of *shake and frown* is regarded as meaning ‘disagree’ although the presence of frown alters the meaning. There is a significant difference between the mean of answers for *shake* versus *shake and frown*.

Dislike. *Frown and tension* appears as the most relevant combination of signals to represent ‘dislike’. But when *shake* is added to *frown and tension*, it alters the meaning. *Frown* alone is sometimes regarded as meaning ‘dislike’ but it is significantly less relevant than *frown and tension*. When displayed on its own, *tension* is also less relevant than the combination *frown and tension*.

Disbelieve. Subjects considered that the combination *tilt and frown* means ‘disbelieve’ (21 answers out of 30) whereas *tilt* on its own is regarded as disbelieve by only 8 subjects. Similarly, *frown* on its own means ‘disbelieve’ for only six subjects. Also, *raise left eyebrow* is regarded by 21 subjects as ‘disbelieve’.

Don’t understand. *Frown* and *tilt and frown* are both associated with the meaning ‘don’t understand’ by 20 subjects. As *tilt* is only given by four subjects one can infer that *frown* is the most relevant signal of the combination. However, when associated to other signals such as *tension* and/or *shake*, *frown* is less regarded as meaning 'don't understand'. Apart from the *frown* signal, *raise left eyebrow* appears as relevant to mean 'don't understand'. It is judged so by 19 out of 30 subjects.

Not interested. For this meaning, two signals seem to be relevant: *eyes roll up* (20 subjects) and *tilt and gaze* (20 subjects). As far as *tilt and gaze* is concerned, it seems it is the combination of both signals that is meaningful since the difference between *tilt and gaze* and *tilt* (13 answers) is significant. Similarly, the difference between *tilt and gaze* and *gaze right down* (13 answers) is also significant.

The results of this test suggest some prototypical signals for most of the meanings. For the positive meanings, 'agree' is signalled by a *nod*; 'accept' is as well. To signal 'like' a smile appears to be the most appropriate signal. A nod associated with a raise of the eyebrows seems to convey 'understand' but only 17 subjects out of 30 thought so. As for 'interested' and 'believe' the experiment did not find prototypical signals. A combination of *smile and raise eyebrows* is a candidate for 'interested'.

For the negative meanings, 'disagree' and 'refuse' are indicated by a head shake; 'dislike' is represented by a *frown and tension* of the lips. A *tilt and frown* as well as a *raise of the left eyebrow* means 'disbelieve' for most of the subjects. The best signal to mean 'don't understand' seems to be a *frown* while *tilt and gaze right down* as well as *eyes roll up* are more relevant for the meaning 'not interested'.

It also appeared that a combination of signals could significantly alter the perceived meaning or that for certain meanings only a composite expression could count as an appropriate signal. For instance, *tension* alone and *frown* alone do not mean 'dislike', but the combination *frown and tension* does. The combination *tilt and frown* means 'disbelieve' whereas *tilt* alone and *frown* alone do not convey this meaning. *Tilt* alone and *gaze right down* alone do not mean 'not interested' as significantly as the combination *tilt and gaze*. Conversely the signal *frown* means 'don't understand' but when the signal *shake* is added, *frown and shake* significantly loses this meaning.

The perceptual experiment aimed to analyse how users interpret context-free backchannel signals displayed by a virtual agent. The result lets one tentatively to assign specific signals to most of the meanings proposed in the test and thus form a start to define a library of prototypes. It remains to see to what extent these form-meaning mappings generalise to other cultures and other contexts. We continue with the description of a similar experiment that investigated the use of vocalisations called affect bursts as backchannels.

3.2 Affect Bursts

Affect bursts are 'very brief, discrete, nonverbal expressions of affect in both face and voice as triggered by clearly identifiable events' (Scherer, 1994, p. 170). Their vocal form ranges from non-phonemic vocalisations such as laughter or a rapid intake of breath, via phonemic vocalisations such as [a] or [m] where prosody and voice quality are crucial to conveying an emotion, to quasi-verbal interjections such as English 'yuck' or 'yippee' for which the segmental form transports the emotional meaning independently of the prosody.

In a study by Schröder et al. (2006) a listening test was carried out to assess the perception of these short nonverbal emotional vocalisations emitted by a listener as feedback to the speaker. The test investigated the use of affect bursts as a means of giving emotional feedback via the backchannel. The acceptability of affect bursts when used as listener feedback seemed to appear to be linked to display rules for emotion expression. While many ratings were similar between Dutch and German listeners, a number of clear differences were found, suggesting language-specific affect bursts.

In a study by Schröder (2003), a range of affect bursts was collected for each of 10 emotions, produced in isolation by German actors. On the basis of phonetic similarity, they were grouped into 24 'affect burst classes', which were classified correctly in a listening test 81% of the time on average. Characterisations of each affect burst class were obtained in terms of the emotion dimensions arousal, valence and power. The distinction between quasi-verbal, language-specific 'affect emblems' and universal 'raw affect bursts', proposed by Scherer (1994), was operationalised in terms of the stability of the segmental form across subjects, which was assessed in a transcription task. This allows one to classify proposed candidates for the status of 'emblem' versus 'raw burst'.

In Schröder et al. (2006) the use of affect bursts as a way for the listener to give emotional feedback was investigated. This is described here.

3.2.1 The Role of Context in Emotion Perception

Context is one of the important factors in the interpretation of expressions. In previous research some important contextual effects were described for the emotional meanings of expressions. Cauldwell (2000) demonstrated that short utterances can be perceived as anger in isolation and as emotionally neutral when perceived in the context in which they were uttered. Interestingly, the perception of anger from the utterance in isolation persisted even after having heard it in context. Similarly, Trouvain (2004) showed that certain kinds of laughter are perceived as sobs in isolation, but as laughs in context. In both cases, the difference in perception was the consequence of *extracting* a vocal expression from its original context. It is unclear whether a similar phenomenon should be expected when a vocalisation which originally was produced in isolation by an actor is inserted into a new context.

Embedding expressive vocalisations into a new context is not a straightforward thing to do, however. Inserting laughs into a speech synthesis context, it was found by Trouvain and Schröder (2004) that most were perceived as inappropriate, with the exception of a very mild laugh. The details of the circumstances under which such an insertion was considered appropriate are not yet clear. In addition, a conversational context may change the *function* of an emotional expressive display. In the case of facial expressions, for instance, Bavelas and Chovil (1997) showed how facial displays of emotion during conversations may not be the result of the emotion felt at the time of speaking but that often they are symbolic parts of messages that are integrated with other communicative signals such as words, intonation and gestures. For instance, a 'surprise' expression may thus be used in a particular context to signal disbelief. Similarly, the interpretation of affect bursts introduced into the conversational backchannel may or may not be interpreted as a comment, a symbolic act rather than the mere expression of an emotion felt. This may influence both the judgements of what is being expressed by the affect burst and the judgements on the appropriateness of the affect burst in this context.

The experiment described in Schröder et al. (2006) addressed the question whether affect bursts can be used by a listener to give emotional feedback to the speaker.

For each of the 10 emotion categories studied by Schröder (2003), 2 affect bursts were selected which were recognised best in isolation; if possible they were chosen from two different affect burst classes. This was possible for all emotions except ‘threat’ and ‘elation’, where both affect bursts had to be selected from the same class. Table 2 lists the original recognition rates of the selected affect bursts along with their respective emotion and affect burst class.

Stimuli were created by embedding each of the 20 selected affect bursts into a neutral speaker sentence. That sentence was deliberately semantically underspecified and spoken in an inexpressive, colloquial way. The sentence was ‘Ja, dann hab’ ich mir gesagt, probierste’s einfach mal ⟨⟨pause⟩⟩ und dann hab’ ich das gemacht!’

Table 2 Recognition results of 20 affect bursts. de = German listeners; nl = Dutch listeners. Ratings of affect bursts in isolation for German listeners taken from Schröder (2003). Acceptability ratings ranged from 0 (very bad) to 100 (very good)

		Recognition (%)					
		Isol.		In context		Acceptability	
Emotion	Burst	de	nl	de	nl	de	nl
Admiration	wow	95	100	97	89	79	70
	boah	95	23	100	11	73	36
Threat	hey1	95	41	70	37	26	23
	hey2	90	19	55	22	26	38
Disgust	buäh	100	69	97	59	53	37
	ih	95	97	90	82	53	45
Elation	ja1	85	90	90	74	51	52
	ja2	70	44	80	40	49	68
Boredom	yawn	95	100	97	96	58	49
	hmm	85	81	86	85	70	51
Relief	sigh	100	100	93	74	46	56
	uff	100	88	90	78	47	45
Startle	int. breath	100	100	100	96	33	34
	ah	90	74	87	48	22	41
Worry	oje	100	34	87	58	62	45
	oh-oh	85	71	97	65	65	45
Contempt	pha	95	81	87	82	35	48
	tse	100	71	87	77	55	50
Anger	growl1	90	81	80	74	37	23
	growl2	80	58	70	48	32	22
Average		92	71	87	65	49	44

(German); ‘Ja, toen zei ik tegen mezelf, probeer het maar een keer ⟨⟨pause⟩⟩ en toen heb ik het gedaan!’ (Dutch); ‘Yeah, then I told myself, why don’t you try it ⟨⟨pause⟩⟩ and then I did it!’ (English translation). In both the German and the Dutch sentence, the pause was 750 ms long. The affect bursts were mixed into the sentence starting at 150 ms into the pause, without modifying the pause duration. In other words, the feedback and the second part of the speaker utterance overlapped for those affect bursts that were longer than 600 ms. All affect bursts were normalised to the same average power as the sentence into which they were embedded. In order to mask the different recording conditions between the speaker sentence and the feedback, a low-intensity white noise (at – 60 dB) was added to the resulting stimuli.

The test was carried out in a web-enabled setup, using the open source tool RatingTest. The 20 stimuli were presented in an automatically randomised order. For each stimulus, subjects answered two questions. In a forced choice setup comparable to the one used by Schröder (2003), they identified the emotion expressed by the listener from a list of 10 categories. In addition, they rated on a continuous scale the question of how well the listener’s interjection fits into the dialogue.

In the German test, 30 subjects participated (15 female; mean age: 24.1 years). And 11 of these took the test in a controlled setting in a quiet office room; the remaining subjects took part in the test via the web. In the Dutch test, 27 subjects participated via the web (5 female; mean age: 24.2 years). A separate group of 32 Dutch listeners also rated the affect bursts in isolation, in order to provide Dutch data comparable to the results in Schröder (2003).

3.2.2 Results

The first observation to make in Table 2 is that the recognition rates for affect bursts in isolation are lower for Dutch listeners than for German listeners. Differences are rather small for the vast majority of bursts; only four bursts that were highly recognised by German listeners are not recognised by Dutch listeners. The two threat bursts were badly recognised, confirming the finding in Schröder (2003) that the threat and anger categories cannot be fully distinguished. Also, Dutch listeners do not seem to make the clear distinction that Germans make between ‘boah’ (expressing admiration) and ‘buäh’ (expressing disgust), leading to a very low recognition for ‘boah’. Similarly low is the recognition of worry ‘oje’, suggesting that in both cases, the language-specific segmental form may be crucial to the emotional meaning.

Regarding the recognition in context, it can be seen from Table 2 that overall recognition rates are slightly lower than for perception in isolation. However, the distribution of recognition rates across categories is very similar to the perception in isolation. One can conclude that the role of context on emotion recognition in this case appears to be very small.

Acceptability ratings showed clear differences between the stimuli, but the pattern is not easy to interpret. One can observe (Table 2) that ratings tend to be consistent within emotion categories. Acceptability was rated very high for admiration (leaving aside the Dutch rating of the ‘boah’ burst not recognised as admiration); moderately high for boredom, worry, elation and relief; moderately low for disgust and contempt; and very low for threat, anger and startle.

Interpretation is not made easier by the inherent ambiguity of the question of ‘good fit’ that the subjects were asked to rate. It may have been interpreted by the subjects as a general appropriateness in the context, as was intended; or one might have found it strange as a reaction to the meaning of the carrier sentence; it may also have been used to indicate technical aspects such as a mismatch between the sound quality of context and burst or the timing of the burst; finally it may have been used to indicate social appropriateness in the given context, in the sense of Ekman’s *display rules*: social norms prescribed by one’s culture as to ‘who can show what emotion to whom, when’ (Ekman, 1977).

Pursuing this issue of social appropriateness, one can attempt to account for the pattern found in terms of display rules. The results can make sense if seen as a cue to display rules whose underlying logic classifies emotions in terms of their being positive or negative and the type of goal they monitor (Castelfranchi, 2000; Poggi and Germani, 2003).

The first display rule seems to point at a general bias against expressing negative emotions. More specifically, the most sanctioned emotions are those linked to goals of aggression (anger and threat), while a somewhat lower sanction holds over negative emotions linked to goals of evaluation (disgust and contempt). Moving up to higher scores, one finds worry, relief and elation, emotions linked to the goal of well-being, and then, even higher, admiration, linked to the evaluation of others. Therefore, a positive bias towards the expression of emotions may hold, first, over emotions that show a positive evaluation of the other (admiration); then positive emotions like elation and relief; and finally over negative emotions like worry. Actually, there is a common feature to elation, relief and worry when expressed after another sentence: They may all be viewed as empathic reactions to the other’s narration.

The experiments described in this section have focussed on how backchannel expressions can express the attitudes of listeners in a conversation rather than at their conversation management functions. From the experiments it appears that the listener responses can have important interpersonal functions. In the context of embodied conversational agents, the relationship between feedback and the effects on the interpersonal relationship has been looked at most closely in the context of rapport. This is discussed in the next section.

4 Agents That Build Rapport

This section presents the Rapport Agent (Gratch et al., 2006b). This agent attempts to create a sense of rapport simply by generating listening feedback based on shallow observable features of a speaker’s bodily movements and speech prosody. We discuss the results of a study that demonstrates the Rapport Agent can produce some of the beneficial social effects associated with rapport. Such agent technology has potential as a powerful and novel methodological tool for uncovering the key factors that influence rapport in face-to-face interactions. It also has potential as a training system to enhance communication skills – for example, to reduce the impact of public speaking anxiety (Pertaub et al., 2001) – or to teach students to recognise specific patterns of nonverbal feedback, such as those that might predict clinical pathologies (Bouhuys and van den Hoofdakker, 1991), those that might cause inter-cultural misunderstandings (Gratch et al., 2006a) or those that arise in the context of deception.

Up to now, only a few systems can condition their listening responses to features of the user's speech, though typically this feedback occurs only after an utterance is complete. For example, Neurobaby analyses speech intonation and uses the extracted features to trigger emotional displays (Tosa, 1993). More recently, Breazeal's Kismet system extracts emotional qualities in the user's speech (Breazeal and Aryananda, 2002). Whenever the speech recogniser detects a pause in the speech, the previous utterance is classified (within 1 or 2) as indicating approval, an attentional bid or a prohibition, soothing or neutral. This recognition feature is combined with Kismet's current emotional state to determine facial expression and head posture. People who interact with Kismet often produce several utterances in succession, thus this approach is sufficient to provide a convincing illusion of real-time feedback.

Only a few systems can interject meaningful nonverbal feedback during another's speech and these methods usually rely on simple acoustic cues. For example, REA will execute a head nod or paraverbal (e.g. 'mm-hum') if the user pauses in mid-utterance (Cassell et al., 1999). Also the Gandalf system produced gaze shifts, back-channel feedback in real time based on the automatic analysis of prosody and gesture input (Thórisson, 1996).

Some work has attempted to extract extra-linguistic features of a speaker's behaviour, but not for the purpose of informing listening behaviours. For example, Brand's voice puppetry work attempts to learn a mapping between acoustic features and facial configurations inciting a virtual puppet to react to the speaker's voice (Brand, 1999).

In all of the cases the feedback by the agent is produced relying on a shallow analysis of some superficial features in the speaker's speech or nonverbal expressions. The feedback that is being produced is mostly intended as showing contact, attention and engagement (Sidner and Lee, 2007), but does not contain much other content. (Jonsdottir et al., 2007, made a first timid attempt to provide more contentful feedback.) The reliance on superficial features seems to be warranted by an experience that most of us have had that it is possible to signal attention by providing feedback even if one is attending only superficially while being preoccupied with other things (Bavelas et al., 2000) – which leads Schegloff (1982) to claim that the term *signal* may not be correct.

> It is worth noting, however, that 'uh huh', 'mm hmm', 'yeah', head nods, and the like *claim* attention and/or understanding, rather than 'showing' it or 'evidencing' it.

Although the feedback produced by listening agents may be based on a shallow analysis, this is not to say that it only has effects on the quality of the process of communication. The feeling of engagement that the feedback is supposed to create will also have an effect on the interpersonal level of communication. Although there is considerable research showing the benefit of such feedback on human to human interaction, there has been almost no research on their impact on human to virtual human rapport (cf. Bailenson and Yee, 2005; Cassell and Thórisson, 1999). In the Rapport Agent, this aspect is being studied in some depth.

Rapport is a crucial factor in establishing successful relationships. Capella (1990) states rapport to be 'one of the central, if not the central, constructs necessary to understanding successful helping relationships and to explaining the development of personal relationships'. It is closely related to some other concepts from social psychology and anthropology, e.g. 'interpersonal sensitivity' (Hall and Bernieri, 2001), 'social glue' (Lakin et al., 2003), 'interactional synchrony' (Bernieri and Rosenthal, 1991), 'mutuality' (Burgoon and Hale, 1987) and empathy (Sonnby-Borgstrom et al., 2003). Tickel-Degnen and Rosenthal (1990) equate rapport with behaviours indicating positive emotions (e.g. head nods or smiles), mutual attentiveness (e.g. mutual gaze) and coordination (e.g. postural mimicry or synchronised movements).[3]

AQ6

That interpersonal rapport is perceptible and is a factor in the success of goal-directed activities is well established in the field of social psychological research. Naive observers will readily make judgements concerning whether participants in dyadic interactions, viewed on video for example, have rapport with one another. A study by Grahe and Bernieri (1999) determined that nonverbal behaviours are more significant than verbal factors in making such judgements. These judgements have been found to correlate reasonably well with the self-assessments of the members of the interacting dyad (Ambady et al., 2000).

Rapport is argued to underlie social engagement (Tatar, 1997), success in teacher–student interactions (Bernieri and Rosenthal 1988), success in negotiations (Drolet and Morris, 2000), improving worker compliance (Cogger, 1982), psychotherapeutic effectiveness (Tsui and Schultz, 1985), improved test performance in classrooms (Fuchs, 1987) and improved quality of child care (Burns, 1984).

Studies have also indicated that rapport can be experimentally induced or disrupted by altering the presence or character of several nonverbal signals (e.g. Bavelas et al., 2000; Drolet and Morris, 2000). Such findings have encouraged the development of embodied conversational agents that can induce rapport through the appropriate generation of nonverbal behavior.

When it comes to creating synthetic agents that simulate human nonverbal behavior, research has focused on half of the equation. Systems emphasise the importance of nonverbal behavior in speech production. Few systems attempt the tight sense-act

[3] See also the chapter by Marinetti et al. 'Emotions in Social Interactions: Unfolding Emotional Experience' in Part I at the beginning of this volume.

loops that seem to underlie rapport and, despite considerable research showing the benefit of such feedback on human to human interaction, few studies have investigated its impact in human to virtual human interaction (cf. Cassell and Thórisson, 1999; Bailenson and Yee, 2005).

4.1 Rapport Agent

The Rapport Agent (Gratch et al., 2006b) was designed at the Institute of Creative Technologies to establish a sense of rapport with a human participant in face-to-face monologs where a human participant tells a story to a silent but attentive listener. In such settings, human listeners can indicate rapport through a variety of non-verbal signals (e.g. nodding, postural mirroring). The fluid, contingent nature of nonverbal behaviour associated with rapport suggests that it could be induced by rapidly responding to a speaker's physical movements. The Rapport Agent attempts to replicate these behaviours through a real-time analysis of the speaker's voice, head motion and body posture, providing rapid nonverbal feedback. The system is inspired by findings that feelings of rapport are correlated with simple contingent behaviours between speaker and listener, including behavioural mimicry (Chartrand and Bargh, 1999) and backchannelling (e.g. nods, see Yngve, 1970). The Rapport Agent uses a vision-based tracking system and signal processing of the speech signal to detect features of the speaker and then uses a set of reactive rules to drive the listening mapping displayed in Table 1. The architecture of the system is displayed in Fig. 2.

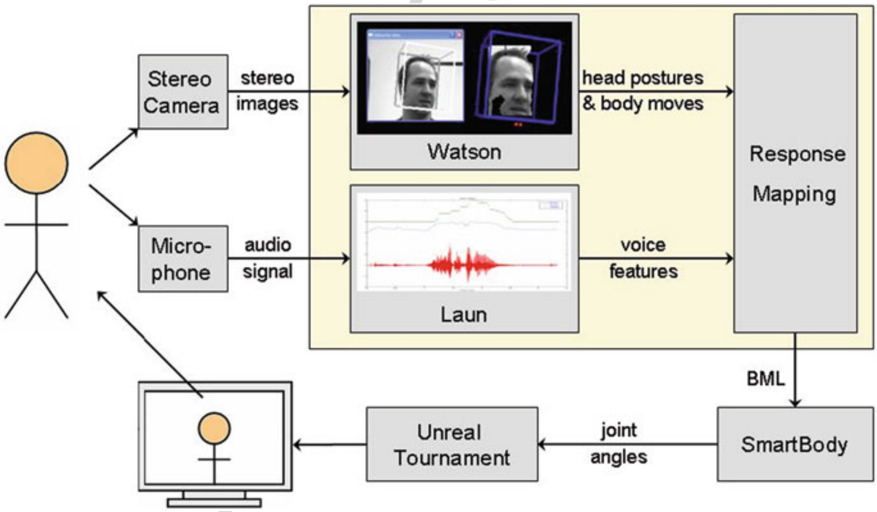

Fig. 2 Rapport Agent Architecture

To produce listening behaviours, the Rapport Agent first collects and analyses the speaker's upper-body movements and voice. For detecting features from the participants' movements, the system detects speaker's head movements. Watson (Morency et al., 2005) uses stereo video to track the participants' head position and orientation and incorporates learned motion classifiers that detect head nods and shakes from a vector of head velocities. Other features are derived from the tracking data. For example, from the head position the Rapport Agent can infer the posture of the spine given that the participant is seated in a fixed chair. Thus, the system detects head gestures (nods, shakes, rolls), posture shifts (lean left or right) and gaze direction.

Acoustic features are derived from properties of the pitch and intensity of the speech signal, using a signal processing package, Laun, developed by Mathieu Morales. Speaker pitch is approximated with the cepstrum of the speech signal (Oppenheim and Schafer, 2004) and processed every 20 ms. Audio artefacts introduced by the motion of the speaker's head are minimised by filtering out low-frequency noise. Speech intensity is derived from amplitude of the signal. Laun detects speech intensity (silent, normal, loud), range (wide, narrow) and backchannel opportunity points (derived using the approach of Ward and Tsukahara (2000).

Recognised speaker features are mapped into listening animations through a set of authorable mapping rules. These animation commands are passed to the SmartBody animation system (Kallmann and Marsella, 2005) using a standardised API (Vilhjalmsson et al., 2007). SmartBody is designed to seamlessly blend animations and procedural behaviours, particularly conversational behaviour. These animations are rendered in the Unreal Tournament™ game engine and displayed to the speaker.

4.2 Evaluation

The social impact of listening feedback has been assessed in a series of formal studies using the Rapport Agent. Some of the key findings are reviewed here (Gratch et al., 2006b, 2007a, b). Studies have conclusively demonstrated that feedback does matter (i.e. different policies for providing listening feedback have a significant impact on speaker fluency, engagement and subjective experience) and that contingency is an important factor (i.e. random feedback gives different results than feedback that is synchronised with features of speaker's behaviour), but that the effects vary depending on individual characteristics of speakers (such as their level of social anxiety).

Interactive virtual agents allow experimenters to carefully manipulate subtle aspects of the feedback and quantify its impact. Studies have contrasted several variants of the Rapport Agent, including a non-responsive agent that displays only random posture shifts, a non-contingent agent that provides the same distribution of feedback as the Rapport Agent but disrupts feedback synchrony (subjects actually see the feedback that was given to a different speaker), an avatar condition that accurately displays the actual movements of a human listener, as well as compared performance with face-to-face interaction.

All studies have involved speakers retelling a recently watched movie (either a funny Sylvester and Tweety cartoon or a serious presentation about sexual harassment in the workplace) to the agent (or a human listener).[4]

Findings show that the presence of listening feedback tends to improve listener performance along several dimensions. When compared with agents that did not provide positive listening feedback (i.e. the unresponsive agent), the Rapport Agent produced more engagement as indexed by the length of stories produced by speakers (Gratch et al., 2006b) and elicited more fluent speech, meaning speakers produced fewer filled pauses, repetitions and broken words (Gratch et al., 2006b, 2007a). One study found that the Rapport Agent could even elicit longer stories than face-to-face interaction between strangers (Gratch et al., 2007b). In general, engagement is positively correlated with the amount of positive feedback, i.e. agents or people that generated more nods tended to elicit longer stories.[5]

Findings also demonstrate that the feedback must be well timed to features of the speaker's behaviour to achieve these beneficial effects, i.e. random feedback is inadequate. When compared with the Rapport Agent or face-to-face interaction, the non-contingent agent produces significantly higher levels of speech disfluency, including far more broken words, repetitions and filled pauses (Gratch et al., 2006b, 2007a, b). This suggests that speakers were distracted by ill-timed feedback, possibly resulting in higher cognitive load. Indeed, subjects rated the non-contingent agent as highly distracting.

Finally, speaker's subjective feeling about the interaction varied with the quantity and quality of feedback, although when compared with observable behaviour (e.g. number of words and disfluencies produced), feelings depend on additional factors, such as their disposition to be anxious in social situations. For example, findings show that subjects that rated high in social anxiety were much more sensitive to non-contingent feedback, reporting higher embarrassment and lower self-perception of performance when compared with less anxious subjects. This suggests the contingency of feedback is especially critical to people who are socially anxious.

Collectively, the findings suggest that virtual agents can achieve some of the elements of rapportful interaction simply by recognising and responding to low-level features of a speaker's non-verbal behaviour. By improving the quality of such feedback, extending its scope to include more features such as gaze and facial expressions and, ultimately, by blending these low-level behaviours with the higher-level semantic understanding more commonly explored by embodied conversational agents, one may be able to realise many of the empirical benefits of rapport on learning and persuasion.

[4] It should be noted that interactions with virtual characters can vary depending on if subjects believe the character is an avatar (controlled by a human) or an agent (controlled by software). In the results we report here, subject were led to believe they were interacting with an avatar to assess the impact of the quality of feedback while holding other factors constant.

[5] It should be noted that listening agents that produced more head nods were also rated as more insincere, arguing for some caution when generating listening feedback.

4.3 Conclusion

In this chapter it was shown how human communication involves a complex synchronisation of actions of multiple participants that are highly connected. Each action calls forth a next one and simultaneously constitutes a reply to a previous one. It is successful or appropriate only in the context of actions that go on in parallel. How actions in human–human communication are intertwined has been studied intensively by linguists, psychologists and sociologists. Creating artificial systems that show the same proficiency in producing behaviours that are equally contingent on the behaviours of human interlocutors is a big challenge. However, it is obvious from studies such as those reported on above that when we want to create virtual agents that we would like to interact with, the agents should be able to at least pretend that they are listening to what we have to say.

Acknowledgments Many of the sections of this chapter are based on previous work of the authors that was done in collaboration with other colleagues. Therefore they wish to acknowledge the big help they got from Sue Duncan, Marion Tellier and Rieks op den Akker.

References

Allwood J (1993) Feedback in second language acquisition. In: Perdue C. (edi) Adult language acquisition. Cross linguistic perspectives. Cambridge University Press, Cambridge, NY, pp 196–232

Allwood J, Cerrato L (2003) A study of gestural feedback expressions. In: Paggio P, Jokinen K, Jonsson A. (eds) First nordic symposium on multimodal communication, Copenhagen, 23–24 September, pp 7–22

Allwood J, Nivre J, Ahlsén E (1993) On the semantics and pragmatics of linguistic feedback. Semantics 9(1)

Ambady N, Bernieri FJ, Richeson JA (2000) Toward a histology of social behavior = Judgment accuracy from thin slices of the behavioral stream. Academic, San Diego, CA, pp 201–271

Austin JA (1962) How to do things with words. Oxford University Press, London

Bailenson JN, Yee N (2005) Digital chameleons = automatic assimilation of nonverbal gestures in immersive virtual environments. Psychol Sci 16:814–819

Bakhtin M (1999) The problem of speech genres. In: Jaworski A, Coupland N. (eds) The discourse reader. Routledge

Bavelas JB, Chovil N (1997) Faces in dialogue. In: Russell J, Fernandez-Dols JM, (ed) The psychology of facial expressions. Cambridge University Press, Cambridge, pp 334–346

Bavelas JB, Coates L, Johnson T (2000) Listeners as co-narrators. J Pers Soc Psychol 79(6): 941–952

Bernieri FJ, Rosenthal R (1991) Interpersonal coordination, behavior matching and interactional synchrony. In: Feldman RS, Rimé B, (ed) Fundamentals of nonverbal behavior. Cambridge University Press, Cambridge

Bevacqua E, Heylen D, Pelachaud C, Tellier M (2007) Facial feedback signals for ECAs. In: Proceedings of AISB'07: artificial and ambient intelligence, Newcastle University, Newcastle upon Tyne, UK, April 2007

Bouhuys AL, van den Hoofdakker RH (1991) The interrelatedness of observed behavior of depressed patients and of a psychiatrist. An ethological study on mutual influence. J Affect Disorders 23:63–74

Brand M (1999) Voice puppetry. In: ACM SIGGRAPH. ACM Press/Addison-Wesley

Breazeal C, Aryananda L (2002) Recognition of affective communicative intent in robot-directed speech. Autono Robots 12:83–104

Burgoon J, Hale J (1987) Validation and measurement of the fundamental themes of relational communication. Commun Monogr 54:19–41

Burns M (1984) Rapport and relationships. The basis of child care. J Child Care 2:47–57

Capella JN (1990) On defining conversational coordination and rapport. Psychol Inquiry 1(4): 303–305

AQ9 Cassell J, Thórisson KR (1999) The power of a nod and a glance, envelope vs. emotional feedback in animated conversational agents. Int J Appl Artif Intell 13(4–5):519–538

Cassell J, Bickmore T, Billinghurst M, Campbell L, Chang K, Vilhjálmsson H, Yan H (1999) Embodiment in conversational interfaces. Rea. In: Conference on human factors in computing systems, Pittsburgh, PA, pp 520–527

Castelfranchi C (2000) Affective appraisal versus cognitive evaluation in social emotions and interactions. In: Paiva A, (ed) Affective interactions. Towards a new generation of computer interfaces. Springer, Berlin

AQ10 Cauldwell R (2000) Where did the anger go? The role of context in interpreting emotion in speech. In: Proceedings of the ISCA workshop on speech and emotion, Northern Ireland, pp 127–131, 2000. URL http://www.qub.ac.uk/en/isca/proceedings

Chartrand TL, Bargh JA (1999) The chameleon effect, the perception-behavior link and social interaction. J Personal Soc Psychol 76(6):893–910

Chovil N (1991) Social determinants of facial displays. J Nonverbal Behav 15:141–154

Clark H, Krych MA (2004) Speaking while monitoring addressees for understanding. J Mem Lang 50:62–81 Clark.Krych.04.pdf

Clark H, Schaefer E (1991) Contributing to discourse. Cogn Sci 13:259–294

Clark HH (1996) Using language. Cambridge University Press, Cambridge

Cogger JW (1982) Are you a skilled interviewer? Personnel J 61:840–843

Condon WS, Ogston WD (1966) Sound film analysis of normal and pathological behavior patterns. J Nerv Dis 143(4):338–347

Drolet AL, Morris MW (2000) Rapport in conflict resolution = accounting for how face-to-face contact fosters mutual cooperation in mixed-motive conflicts. Exp Soc Psych 36:26–50

Duncan S, Fiske DW (1977) Face-to-face Interaction. Erlbaum, Hillsdale NJ

Duncan SD, Niederehe G (1974) On signalling that its your turn to speak. J Exp Soc Psychol 10:234–47

Ekman P (1977) Biological and cultural contributions to body and facial movement. In: Blacking J, (ed) The anthropology of the body. Academic, London, pp 39–84

Ekman P (1979) About brows: emotional and conversational signals. In: Cranach von M, Foppa K, Lepenies W, Ploog D, (ed) Human ethology. Cambridge University Press/Editions de la Maison des Sciences de l'Homme, Cambridge, pp 169–202

Fuchs D (1987) Examiner familiarity effects on test performance, implications for training and practice. Top Early Child Spec Educ 7:90–104

Goffman E (1976) Replies and responses. Lang Soc 5(3):2257–313

Goodwin C (1984) Notes on story structure and the organization of participation. In: Atkinson MJ, Heritage J, (ed) Structures of social action. studies in conversation analysis. Cambridge University Press, Cambridge, pp 225–246

Grahe JE, Bernieri FJ (1999) The importance of nonverbal cues in judging rapport. J Nonverbal Behav 23:253–269

AQ11 Gratch J, Okhmatovskaia A, Duncan S (2006a) Virtual humans for the study of rapport in cross cultural settings. In: 25th army science conference, Orlando, FL, 2006a

Gratch J, Okhmatovskaia A, Lamothe F, Marsella S, Morales M, van der Werf R, Morency LP (2006b) Virtual rapport. In: 6th international conference on intelligent virtual agents, Springer, Berlin Marina del Rey, CA

Gratch J, Wang N, Gerten J, Fast E (2007a) Creating rapport with virtual agents. In: 7th international conference on intelligent virtual agents, Paris, France, 2007a

Gratch J, Wang N, Okhmatovskaia A, Lamothe F, Morales M, van der Werf R, Morency L-P (2007b) Can virtual humans be more engaging than real ones? In 12th international conference on human-computer interaction, Beijing, China

Grice HP (1975a) Logic and conversation. In: Cole P, Morgan JL, (ed) Syntax and semantics: Vol 3: speech acts. Academic, San Diego, CA, pp 41–58
Grice HP (1975b) Meaning. Phil Rev 66(3):377–388
Gumperz J (1982) Discourse strategies. Cambridge University Press, Cambridge, England
Hadar U, Steiner TJ, Rose CF (1985) Head movement during listening turns in conversation. J Nonverbal Behav 9(4):214–228
Hall J, Bernieri F. (ed) Interpersonal sensitivity. LEA, Mahwah, NJ
AQ12 Heylen D (2007) Multimodal backchannel generation for conversational agents. In: Proceedings of the workshop on multimodal output generation (MOG 2007), University of Twente, 2007. CTIT Series, p 8192
AQ13 Heylen D, Bevacqua E, Tellier M, Pelachaud C (2007a) Searching for prototypical facial feedback signals. In: IVA, 2007a pp 147–153
Heylen D, Nijholt A, Poel M (2007b) Generating nonverbal signals for a sensitive artificial listener. In: Esposito A, Faunder-Zanny M, Keller E, Marinaro M. (ed) Verbal and nonverbal communication behaviours, Lecture notes in computer science. Springer, Berlin, pp 264–274
Jefferson G (1984) Notes on a systematic deployment of the acknowledgement tokens ‘yeah’ and ‘mm hm’. Papers in Linguistics, 17:197–206
AQ14 Jonsdottir GR, Gratch J, Fast E, Thórisson KR (2007) Fluid semantic back-channel feedback in dialogue: challenges and progress. In: IVA, 2007 pp 154–160
Kallmann M, Marsella S (2005) Hierarchical motion controllers for real-time autonomous virtual humans. In: Intelligent virtual agents, Kos, Greece, 2005. Springer, Berlin
Kendon A (1970) Movement coordination in social interaction: some examples described. Acta Psychol 32:100–125
Kendon A (1967) Some functions of gaze direction in social interaction. Acta Psychol 26:22–63
Kraut RE, Lewis SH, Swezey LW (1982) Listener responsiveness and the coordination of conversation. J Pers Soc Psychol 43(4):718–731
LaFrance M (1979) Nonverbal synchrony and rapport: analysis by the cross-lag panel technique. Soc Psychol Quart 42(1):66–70
Lafrance M, Ickes W (1981) Posture mirroring and interactional involvement: sex and sex typing effects. J nonverb behav 5:139–154
Lakin JL, Jefferis VA, Cheng CM, Chartrand TL (2003) Chameleon effect as social glue, evidence for the evolutionary significance of nonconscious mimicry. J Nonverb Behav 27(3):145–162
McClave EZ (2000) Linguistic functions of head movements in the context of speech. Journal of Pragmatics, 32:855–878
AQ15 Morency L-P, Sidner C, Lee C, Darrell T (2005) Contextual recognition of head gestures. In: 7th international conference on multimodal interactions, Torento, Italy, 2005, pp 18–24
Oppenheim AV, Schafer RW (2004) From frequency to quefrency. A history of the cepstrum. IEEE Signal Process Mag September:95–106
Pelachaud C, Bilvi M (2003) Computational model of believable conversational agents. In: Huget M-P (ed) Communication in multiagent systems, vol 2650 of Lecture notes in computer science. Springer, Berlin, pp 300–317
Pertaub D-P, Slater M, Barker C (2001) An experiment on public speaking anxiety in response to three different types of virtual audience. Presence Teleoperators and Virtual Environ 11(1): 68–78
Poggi I (2007) Minds, hands, face and body. Weidler Buchverlag, Berlin
Poggi I, Germani M (2003) Emotions at work. In: International conference on human aspects of advanced manufacturing: agility and hybrid automation, Rome, Italy, 2003, pp 461–468
AQ16 Rosenfeld HM (1987) Conversational control functions of nonverbal behavior. In: Nonverbal behavior and communication. Lawrence Erlbaum Associates, Hillsdale NY, pp 563–601
Rosenfeld HM, Hancks M (1980) The nonverbal context of verbal listener responses. In: Key MR (ed) The relationship of verbal and nonverbal communication. Mouton, The Hague, pp 193–206
Sacks H, Schegloff EA, Jefferson G (1974) A simplest systematics for the organization of turn-taking for conversation. Language 50:696–735

Schegloff EA (1982) Discourse as interactional achievement: some uses of "uh huh" and other things that come between sentences. In: Tannen D. (ed) Analyzing discourse, text, and talk. Georgetown University Press, Washington, DC, pp 71–93

Schegloff EA, Sacks H (1973) Opening up closings. Semiotica 8:289–327

Scherer K (1994) Affect bursts. In: van Goozen SHM, van de Poll NE, Sergeant JA. (ed) Emotions: Essays on emotion theory. Lawrence Erlbaum, Hillsdale, NJ, pp 161–193

Schröder M (2003) Experimental study of affect bursts. Speech Commun Special Issue Speech a Emot 40(1–2):99–116

Schröder M, Heylen D, Poggi I (2006) Perception of non-verbal emotional listener feedback. In: Proceedings of speech prosody 2006, Dresden, Germany, 2006

Searle JR (1969) Speech acts: an essay in the philosophy of language. Cambridge University Press, Cambridge

Sidner CL, Lee C (2007) Attentional gestures in dialogues between people and robots. In: Nishida T (ed) Conversational informatics. Wiley, New York, NY

Sonnby-Borgstrom M, Jonsson P, Svensson O (2003) Emotional empathy as related to mimicry reactions at different levels of information processing. J Nonverb Behav 27(1):3–23

Tatar D (1997) Social and personal consequences of a preoccupied listener. PhD thesis, Department of Psychology, Stanford University, Stanford, CA

Thórisson KR (1996) Communicative Humanoids: a computational model of psycho-social dialogue skills. PhD thesis, Massachusetts Institute of Technology

Tickle-Degnen L, Rosenthal R (1990) The nature of rapport and its nonverbal correlates. Psychol Inquiry 1(4):285–293

Tosa N (1993) Neurobaby. ACM SIGGRAPH, pp 212–213

Trouvain J (2004) Non-verbal vocalisations – the case of laughter. Paper presented at Evolution of Language: Fifth International Conference, 2004. Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany, 31 March – 3 April 2004

Trouvain J, Schröder M (2004) How (not) to add laughter to synthetic speech. In: Proceedings of Workshop on affective dialogue systems, Kloster Irsee, Germany, 2004 pp 229–232

Tsui P, Schultz GL (1985) Failure of rapport. Why psychotherapeutic engagement fails in the treatment of Asian clients. Am J Orthopsychiatry, 55:561–569

Vilhjalmsson H, Cantelmo N, Cassell J, Chafai NE, Kipp M, Kopp S, Mancini M, Marsella S, Marshall AN, Pelachaud C, Ruttkay ZS, Thorisson KR, van Welbergen H, van der Werf R (2007) The behavior markup language, recent developments and challenges. In: International conference on intelligent virtual agents, Paris, France, 2007. Springer, Berlin

Ward N, Tsukahara W (2000) Prosodic features which cue back-channel responses in English and Japanese. J Pragmatics 23:1177–1207

Yngve VH (1970) On getting a word in edgewise. In: Papers from the sixth regional meeting of the Chicago Linguistic Society. Chicago Linguistic Society, Chicago, IL, pp 567–577