Mutually Coordinated Anticipatory Multimodal Interaction

Anton Nijholt, Dennis Reidsma, Herwin van Welbergen,
Rieks op den Akker, and Zsofia Ruttkay

Human Media Interaction Group (HMI)
Department of Computer Science, University of Twente
The Netherlands
{anijholt,dennisr,welberge,infrieks,zsofi}@cs.utwente.nl

Abstract. We introduce our research on anticipatory and coordinated interaction between a virtual human and a human partner. Rather than adhering to the turn taking paradigm, we choose to investigate interaction where there is simultaneous expressive behavior by the human interlocutor and a humanoid. Various applications in which we can study and specify such behavior, in particular behavior that requires synchronization based on predictions from performance and perception, are presented. Some observations concerning the role of predictions in conversations are presented and architectural consequences for the design of virtual humans are drawn.

1 Introduction

Virtual humans are based on implementations of models of human (expressive) behavior; models that are used to drive the interaction between a virtual human and human 'user'. Human behavior can not completely be understood as the execution of a preconceived program, a set of conditional rules, the application of which depends on the classification of observable events according to a number of preformatted classification schemes. These are the categories of the designer who has a complete specification of the motives that agents drive, the goals they have and the means they can use to realize these goals. On the contrary, the goal of a human activity is the realization of the person 'self'. How the actions become realized depends on the actions of other agents. Interaction is emergent, not the result of planned actions but something that simply shows up as a result of many synchronous activities as well as what has been established before. Thus there is a tension between the perspective of the designer of synthetic humanoid interactive characters and the creative emergent behavior in which humans realize themselves through interaction as social beings. A good example of this tension between emergent human conversational behavior and the design of interactive computers can be found in the discussions about the topic of turn-taking for virtual humans. People tend to talk 'in turns', but they do talk simultaneously as well. However, developers of virtual humans and conversational systems through the years have mostly turned to the turn-taking framework originally set out by Sacks et al. [38]. This work has been used as a normative description of

A. Esposito et al. (Eds.): HH and HM Interaction 2007, LNAI 5042, pp. 70–89, 2008.

coordination principles in conversation (even though it has been criticized as such by other theorists [12,29]): one speaker talks at a time, speakers take turns, a next speaker's turn commonly follows the previous turn with no gap and no overlap, etc. In dialogue systems this leads to a clear pipeline ordering of modules. First the interlocutor's utterance is perceived through the system's sensors (microphone, keyboard, etc), then the utterance is interpreted, next deliberation takes place in the system about the appropriate (re)action to be chosen, and finally the reaction of the system is produced. Although the framework has been criticized in many ways it is still the most used paradigm for conversational systems.

From the 'turn taking' theoretical viewpoint, feedback or backchannel [50] is seen as standing slightly outside the main conversation, helping regulate the turn taking process but not actually part of the main conversational exchange. Furthermore, since the turn taking paradigm privileges speech, the continuously ongoing non-verbal expressions of speakers and listeners cannot readily be given a place in turn taking patterns. It is therefore perhaps not surprising that the people building systems dealing with verbal and non-verbal feedback from listeners were the first to turn to other models of interaction, less embedded in the turn-taking framework. In the next section we will see how frequently speaker overlap occurs in four participant meetings.

In the research reported here we want to put aside the concept of prescriptive turn-taking models entirely. We envision conversational systems where two-way simultaneous interaction is not limited to a speaker speaking and a listener providing some feedback. Rather, we want to see perception and production occurring simultaneously on both sides of the conversation. A major theme then is that of alignment and synchronization. That is, simultaneous interaction (as opposed to sequential pipe-line interaction) needs a much tighter temporal coordination between the expressions of a virtual human on the one hand and the perception of the expressions of the human interlocutor on the other hand. We will discuss the need for anticipatory and predictive models and their place in the interaction. The generation of multi-modal expressions in virtual humans not only involves the planning of their content on a conceptual level, but also those expressions need to be executed using the appropriate modalities, in coordination with the conversational partner. Hence, we need to look at planning and re-planning problems and at the low level animation consequences for virtual humans on the production side, and at types of perception needed for turn-free two-way interaction.

Our research on simultaneous interaction is meant to become integrated in our Sensitive Artificial Listener (SAL) system [17]. Our inspiration comes also from related research on three applications where there is continuous interaction between user and system (virtual human): a dancer, a virtual music conductor and a physio-trainer. Each of those applications consists of an interactive virtual human capable of observing the user through sensors and expressing itself verbally and/or nonverbally. The applications investigate different aspects of simultaneous two-way interaction. The development as well as the evaluation of these applications leads to observations and questions about the design of more natural interaction with virtual humans and about the models needed for such. This paper presents different insights as well as open questions raised while working on the applications. Furthermore, our observations are confronted with related work in virtual human research as well as with relevant literature from linguistics and social psychology.

Section 2 of this paper is devoted to observations on synchronization in conversation. We discuss the useful and well-known turn taking model of Sacks et al. [38], but also comment on its shortcomings, in particular its lack of attention for the simultaneous expressive behavior of speaker and listener. In section 3 we discuss the aforementioned applications of an interactive dancer, an interactive conductor and an interactive trainer. In these applications we can speak of 'anticipatory synchronization' in the interaction, where behavioral timing is guided by music or a rhythm that has to be followed. In section 4 we present a preliminary investigation whether this view of anticipatory synchronization can also be recognized and explored in speech conversations. In section 5 we discuss the architectural consequences of anticipatory synchronization for virtual humans. Again, the initial ideas for an architecture that allows the generation of coordinated behavior started with the three applications mentioned earlier. In this section we discuss the Behavioral Markup Language (BML, [45]) for specifying synchronization of behaviors, including synchronization with external events. Some conclusions are drawn in the final section.

2 Synchronization and Interaction in Conversation

In games, formal social events, business meetings, in traffic, the notion of turn abounds and turns are pre-allocated, i.e. there is a protocol or system of rules that prescribes who may take turn when. In casual conversations speaker turns are not pre-allocated. Interactants "locally manage", that is on a turn by turn basis, who will be the next speaker. In their ground breaking paper Sacks, Schegloff and Jefferson (SSJ) propose a model that describes how turn taking is performed [38]. Such a model should clarify a number of phenomena that anyone can observe in casual conversations. Two of these observations are: 1) time between two adjacent speaker turns is usually small, and 2) speaker overlap is uncommon. According to the turn taking rules of the SSJ system the current speaker decides who will take turn, for example by addressing a question to a selected addressee, but if he doesn't listeners self-select to take turn. Important notion in SSJs system is *turn constructional unit* (TCU) and the related *transitional relevant place* (TRP). A listener will not take over turn at any arbitrary time; he will wait until a TCU has been completed by the speaker. A TCU is a semantical or informational complete phrase: a listener usually will not interrupt a speaker in the middle of such a TCU. TCUs are often prosodically marked phrases ending with a clear rising or lowering of pitch and energy. What is important in the light of the second observation (i.e. short time lags in between two speaker turns) is that TCU do project their ending: listeners recognize a TCU and are thereby often able to predict how long it takes before the TCU has been completely produced by the speaker. The attentive listener actively participates in formulating the ideas that he has recognized and that the speaker is uttering.

In her essay "The Paris Years" Carol Sanders discusses the way a listener identifies the relevant meaning of a word [39]. It both involves, according to Sanders, the concept of the listener as the mirror image of the speaker and as an active part of the 'circuit du langue'. She quotes the French linguist Michel Bréal (1897):

> "It is not even necessary to suppress the other meanings of the word: these meanings do not impinge on us ... for the association of ideas is based on the

> sense of things, not on their sound. What we say about speakers is no less true about listeners. These are in the same situation: their thought run along with or precedes that of their interlocutor. They speak inwardly at the same time as we do: so they are no more likely than we are to be troubled by related meanings dormant in the depth of their minds."

The observation that listeners simultaneously with speakers 'speak', i.e. phrase the idea that is jointly worked out, is basic for understanding the shortness of response times and the shortness or even lack of gaps in turn transitions. Some people have understood SSJs theory to say that a conversation can only be successful if the participants obey the turn-taking rules as described by the model. Others have pointed at the fact that the participants themselves decide how a conversation enrolls. Both Cowley [12] and O'Connell et al. [29] criticize the turn-taking tradition for mistaking the *descriptive* paradigm for a *prescriptive* model. Coates suggests that in some contexts a 'free-for-all' metaphor, where everybody equally can contribute to the conversation at all times, is more appropriate. Then, overlapping speech is a signal of participant's active engagement rather than a signal of conversational malfunction [11]. Bavelas et al. [4] also give a more active role to the listener, extending beyond signaling attention and understanding over an underprivileged back channel to contributing pertinent content to a narrative through specific listener responses (other authors use the term content feedback).

Part of SSJs turn-taking systems concerns repair, describing how interlocutors operate together when they take turn simultaneously. We don't go into the details of this procedure, but note that simultaneous talk is not always considered problematic. Some people are able to listen and talk at the same time; or, for some people it sometimes doesn't matter what others have to say; why shouldn't they speak simultaneously? It is often hard to draw the line between interaction and parts in a conversation where there is no 'real' interaction, in particular because the refusal to interact (turning the back) is also a sign send out to be picked up by others in interaction. According to SSJ politeness, familiarity, and affect are important factors that determine how a conversation emerges. It will be clear that if participants build on the previous contribution they have to wait until the previous speaker is ready (you can't answer a question if you don't know the question). But if you already have understood the question before the speaker has finished the formulation waiting is more a matter of politeness, than a conceptual necessity.

Another comment on SSJ is that it restricts conversational actions to verbal communicative acts only. Some critics say that we should take non-verbal activities into account. If we do, we see that listeners are continuously and simultaneously active along with the speaker; they show behavior that communicates how they receive, and uptake the message being conveyed by the speaker. If for example someone is asking a question and sees during his speaking that the addressee looks puzzled, he will elaborate his question before yielding turn to the addressed person, while if the addressee looks eager to answer the question that he already grasped before the speaker finished his interrogative, the speaker will pass the turn and allow the addressed partner as soon as possible. This is not the place to discuss whether this is a valid critic on SSJs theory and model. What is important is that the notion of turn is highly problematic, in the sense that there is no computational model that for once and for all situations (even not for a restricted type of interactions, like casual conversations) tells us what is a turn and what is not.

Several systems have been implemented by different groups that incorporate different types of backchannelling behavior, which by definition involves simultaneous expressive behavior from a 'speaker' and a 'listener', often using non-verbal signals. In all of those systems, backchannel expressions given by the virtual human are defined and implemented in a 'reactive' way, i.e., upon perceiving certain features in the expressions of the speaker, the system will react directly with certain nonverbal feedback expressions [43,44,20,46,25,17].

Another way of looking at temporal aspects of coordination between humans in a conversation is by looking at a corpus of recordings of small group discussions. The fact that temporal coordination plays an important role in conversation can readily be seen by visualizing some aspects of the interaction, for example looking at the dynamics of the energy level in the sound produced by each of the speakers along the time axis.

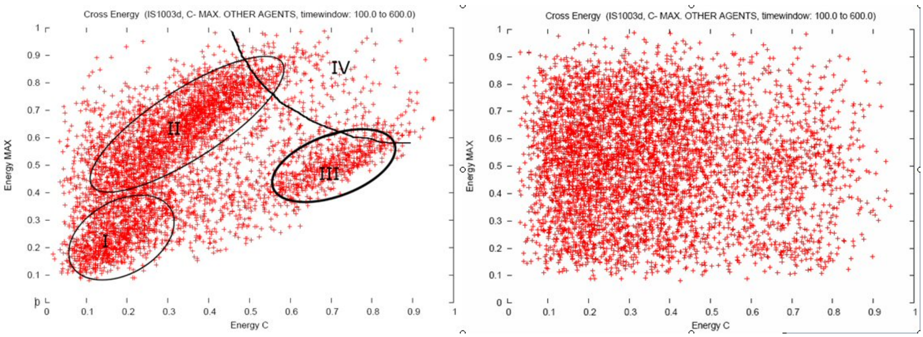


Fig. 1. Temporal coordination between one participant in a meeting and the others, visualized through the relation between energy levels in the audio channels

Figure 1 shows cross energy of audio channels in a four participant face-to-face meeting recorded in the AMI corpus [8]. Both graphs show the relation between the energy in the audio channel of participant C and the maximum energy in the other three audio channels. The energy levels on the X-axis and on the Y-axis in the leftmost graph are clearly correlated; every point in that graph is a point in time in the meeting. Area I is where both energy levels are low (no speech), area II is where C doesn't speak, but someone else speaks, area III is where C speaks and others don't, area IV is where C's talk overlaps with others' talk.

Thus, two things are clear from this graphical representation.

- There is overlap in talk (area IV)
- There are clear patterns of coordinated behavior shown in the energy levels of talk in interaction (see the clusters in areas I, II en III).

The rightmost graph was obtained by shifting all energy values in the audio channel of participant C a small amount in time. The result is a graph that shows no structure anymore.

To make the structure of the temporal coordination between participants in the meeting even clearer, Figure 2 shows a similar graph for two individual participants, namely participant C and D. The number of words spoken by each of the participants ranges from 400 to 450 in the recorded time frame of 400 sec. The dark cloud in the lower left corner (indicating low energy on both channels at the same time) of the graph in Figure 2 results from periods of no speech. Audio is recorded by means of lapel microphones.

If we look at the pictures in Figure 1 we see that the left one shows more structure than the other. This structure reflects the interaction between the participants in the meeting. The sounds produced by them are not completely independent, it shows some pattern, a pattern that recurs in different time frames, and also in different meetings, with different participants. It seems like a pattern that reveals something of the general idea of people having a conversation in a meeting.

As Basu [3] also shows, by making similar pictures of ‘pseudo interactions’ (similar as the one shown in right side of Figure 1), these patterns aren’t a coincidence, but caused by the coordinated actions between different speakers in interaction. Basu uses this fact to find in a set of mixed audio streams those pairs of streams that form the two halves of one and the same conversation.

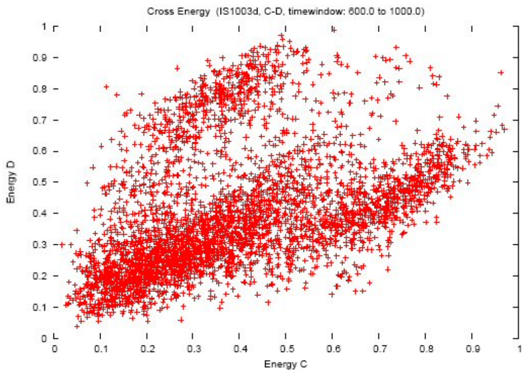


Fig. 2. Temporal coordination between two individual participants in a meeting, visualized through the relation between energy levels in the audio channels

An example of coordinated action is the correlation of gaze behavior and speaking behavior in backchannelling or in addressing. We may say that both ‘behaviors’, gaze and speech, are in fact part of the same activity; moreover they often go along with other ‘behaviors’, such as body movements towards the speaker or addressed person. Head and eye movements are highly dependent, because they participate in the same behavior.

An other illustration showing synchronization between nonverbal activities of participants having a conversation is based on the work of Ramseyer and Tschacher [31] (see also section 4).

Two video streams are recorded from the interaction between two persons: in our case both close-up videos showing head movements of the persons. In each of the video the amount of movements is measured as a function of time, by a simple *image difference* computation. The correlation between the functions is then computed. Since persons do not move their heads exactly at the same time, but with a short delay in response to each other, this computation is done by comparing each window of

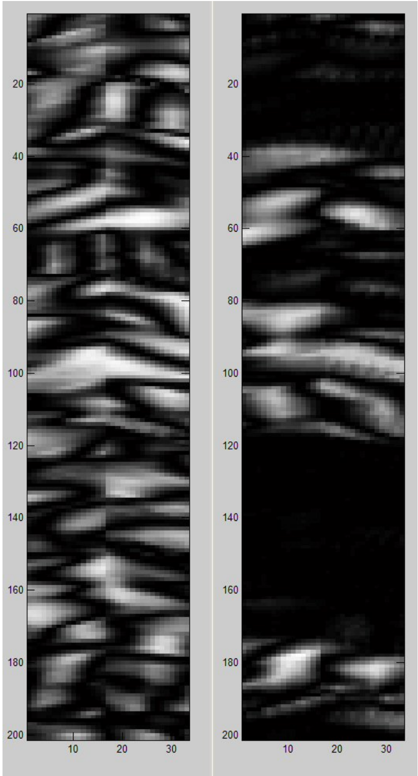


Fig. 3. Visualization of coordinated head-movements of two persons in interaction

person A with all windows of person B that are within a timeframe of -2 and +2 secs of the time of the window of A. We have generated these from data of a fragment of 200 windows (8 sec) in a conversation recorded in the SAL project [17].

In Figure 3 the vertical axis is the time axis and time runs from top to bottom, measured in window units. Each window is 4 secs; with 0.4 sec between two windows. The horizontal axis contains the 'time lag' between two windows (one of A, one of person B) for which the correlation has been computed, from -2 sec to +2 sec. The brighter the point, the stronger this correlation.

The left part of Figure 3 shows the graph for an interaction between persons A and B in the second part of a conversation. The right graphics is made up by cross-correlating two pieces of data of persons A and B from different parts of the conversation: data from A from 8 sec in the first part of a conversation is compared with data of person B from the second part of the conversation.

The conclusion that we can draw from these pictures is that on a local level there exists a meaningful temporal correlation between the movements made by the partners in interaction. Thus, these correlations can be used as cues for detecting interaction.

3 Mutually Coordinated Multi Modal Interaction: 3 Examples

We are building applications in which perception and production are parallel rather than sequential processes. As we have seen above, this needs a certain amount of coordination involving temporal aspects. The applications have all been described in more detail elsewhere - in this section we will shortly summarize them, stressing the aspects of mutual coordination between a virtual human and interlocutor.

3.1 Interactive Virtual Dancer

In a recent application built at HMI, a virtual dancer invites a real partner to dance with her [34]. The Virtual Dancer dances together with a human 'user', aligning its motion to the beat in the music input (see Figure 4). The system observes the movements of the human partner by using a dance pad to register feet activity and the computer vision system to gain information about arm and body movements. By responding to the way the human user is dancing, the virtual dancer implicitly invites the user to react to her as well. At any point in time, the virtual human in this application is both expressing herself (dancing to the beat), and perceiving the user's style of dancing. When the virtual dancer is in a 'following' mode she will break of dancing moves when they no longer fit with the user's style and continue with better fitting moves. When in 'leading' mode she may introduce dance moves with a completely new style in order to evoke reactions from the user.

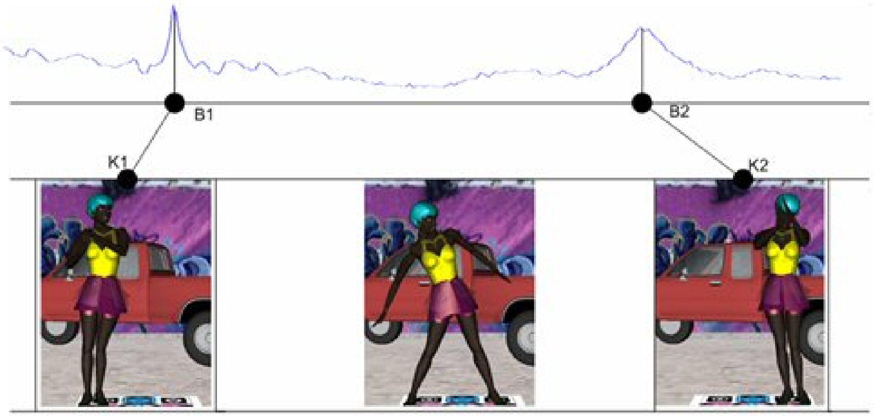


Fig. 4. An interactive virtual dancer dancing to the beat of the music

3.2 Interactive Virtual Conductor

We have designed and implemented a virtual conductor [6] that is capable of leading, and reacting to, live musicians in real time (see Figure 5). The conductor possesses knowledge of the music to be conducted, and it is able to translate this knowledge to gestures and to produce these gestures. The conductor extracts features through audio processing algorithms as the music is played and reacts to them, based on information of the knowledge of the score. The reactions are tailored to elicit the desired response from the musicians.

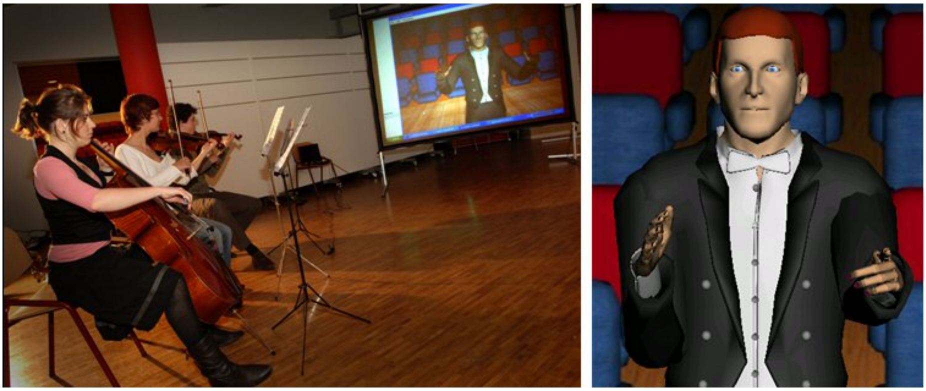


Fig. 5. An interactive virtual conductor conducting a human orchestra

Clearly, if an ensemble is playing too slow or too fast, a (human) conductor should lead them back to the correct tempo. He can choose to lead strictly or more leniently, but completely ignoring the musicians' tempo and conducting like a metronome set at the right tempo will not work. A conductor must incorporate some sense of the actual tempo at which the musicians play in his conducting, or else he will lose control. If the musicians play too slow, the virtual conductor will conduct a little bit faster than they are playing. When the musicians follow, it will conduct faster yet, till the correct tempo is reached again. In order to do this, the Virtual Conductor continuously makes a prediction of how the musicians will be playing in the next few beats, in order to coordinate its conducting behavior to their music.

3.3 Interactive Virtual Trainer

The scenario of the Reactive Virtual Trainer (RVT) describes a virtual human capable of presenting physical exercises that are to be performed by a human, monitoring the user and providing feedback [37]. The reactivity of the RVT is manifested in natural language comments, readjusting the tempo, pointing out mistakes or rescheduling the exercises. Such exercises can be performed at the beat of a user's favorite music. Exercises describe a mix of behaviors on different modalities, including exercise movement, sound (such as clapping, feet tapping), speech and music. This scenario is similar in certain ways to the virtual conductor. The RVT can do the exercises along with the user, adapting its tempo to the performance of the user, or attempting to lead the user when his/her tempo is lagging.

4 Mutual Coordination: Anticipatory Synchronization

In all three applications presented above, it turned out that behavior expressions had to be synchronized to predictions of perception of the environment (dancer and music) or the interlocutor's behavior (trainer, conductor). This is a type of behavior coordination that has hardly been addressed before. On first sight, it seems to be quite specific of the applications. In this section we survey some literature, and look at data in our corpora, to find out whether such 'anticipatory synchronization' only makes sense in the context of these specific applications, or whether there are underlying issues that make anticipatory synchronization a more generally applicable issue for virtual humans.

To start with, consider some literature about minimal reaction times for vocal responses for humans: Wilson and Wilson [49] say that vocal reaction time for highly primed subjects is 200 msec (citing Izdebski and Shipp [18], for a vocal 'react as fast as possible' task without any cognitive processing). Goodrich et al. [15] give minimal reaction times between 200 and 500 msec in a complex task with distraction. Slowiaczek [40] gives reaction times of 700-800 msec for a task that involves lexical lookup. To a certain extent, when more cognitive processing is required, minimal reaction times become larger.

Now consider some literature about the time scale on which certain communicative behaviors actually occur. For example, Nagaoka et al. [27] describe converging response latencies for some dyads in the 300-600 msec range; Jonsdottir et al. [20] show that content feedback comes within 500-1000 msec; Ward and Tsukahara [46] says that envelope feedback occurs in the range of 350 msec after an utterance; Cowley [12] describes some conversation timing effects within 300 msec and shorter range; Wilson and Wilson [49] present data on gaps between utterances where many are below 200 or even 100 msec, which is below the absolute physical reaction time for highly primed subjects.

Given such results it seems reasonable to suggest that the timing of some of the effects described above is in a range shorter than possible purely on reaction times. This observation also underlies the idea of, for example, projectability of utterances in the turn taking theory of Sacks et al. [38]. Humans anticipate the timing of expressions of their interlocutor in order to match their responses to it.

But what use is it to implement such effects of anticipatory synchronization in conversational virtual humans? Literature such as the work of Crown [13], Ramseyer and Tschacher [31] or Nagaoka et al. [27] on interactional synchrony or coordinated interpersonal timing in communication suggests that being able to coordinate one's actions in an anticipatory manner to those of one's interlocutor relates to a positive evaluation of the conversation partner and of the (effectiveness of) the interaction. Crown [13] for example relates interpersonal timing to affective relation in dyads, and concludes that a 'like/dislike/unacquainted' condition has a strong relation with interpersonal timing. Ramseyer and Tschacher [32] were the first to perform a large scale quantitative study of synchrony in a psychotherapeutic context; using the analysis developed by Boker et al. [5] they found a clear correlation between synchrony and certain positive therapy outcomes in a data set of 125 therapy sessions by 80 dyads. A good overview of themes and topics related to synchrony can be found in the survey of Nagaoka et al. [27]. Besides pointing the reader to a large amount of earlier work they discuss experiments with rhythmic entrainment (convergence of latencies between utterance and response, for speakerA/ListenerB and speakerB/listenerA transitions), showing how dynamics and alignment are an important elements of synchrony tendency. This in turn is important for conveying rapport and empathy, promotion of understanding emotion and making you assessed positively by the other. In summary, it appears that coordinated interpersonal timing can convey info about mental state and help influencing/persuading the other in some sense.

So far we have only discussed the precise timing in human-human interaction. But why do we need such precise timing relations in human-computer interaction? In other words, what happens if precise timing in an interaction is disrupted? Such a situation occurs frequently in video conferences, where speech and video is delayed by the transmission over a network. Such delays in transmission disrupt turn taking mechanisms [14]. This causes audio collisions and a reduction of interactivity: the turns are longer and backchannel feedback occurs less. Even when a timing disruption goes unnoticed it can have social impact. In [33], it is shown that a delay between audio and video causes users to evaluate a speaker as less interesting, less pleasant, less influential, more agitated and less successful in their delivery, even if they did not notice the asynchrony itself.

More specifically in the context of human computer interaction, we do know that at least to a certain extent interactional synchrony also works for human-VH interaction. For example, Suzuki et al. [41], working on prosody, say that echoic humming mimicry has a positive influence on affective perception of the conversational partner, even if that partner is a computer. Bailenson and Yee [2] also specifically addresses the dynamics of the movement: interactional synchrony, in the form of mimicry (repeat head movements of partner after 4 secs) is effective for VHs to be more persuasive and effective. This is all not very surprising, as Reeves and Nass [33] already showed that this type of aspects in human-human communication transfer to human media communication.

More on the topic of timing, Robins et al. [36], working with robots rather than virtual humans, conclude qualitatively from an exploratory study about “Rhythm, kinesics, body motion and timing” that “[...] responding with appropriate timing so as to mesh with the timing of human actions encourages sustained interaction” and “Robot-human temporal interaction kinesics will eventually need to be studied deeply in order to put this dimension within the purview of HRI designers”.

However, modeling and implementing anticipatory synchronization in conversational virtual humans is easier said than done. In the applications presented above the music, or the fitness exercise, defines a rhythmic structure that can be taken as starting point for the prediction of the behavioral timing of the interlocutor. But in normal conversation one does not match ones behavior to that of the interlocutor through the mediation of music or another externally defined rhythm. When we want to extend these concepts to conversations with a virtual human we should explore the vast amount of literature on investigations into the rhythmic nature of speech. Although it is not trivial to find a rhythmic organization that can be used for predicting e.g. turn endings [7] or other practical uses [21]. So, if we want to have anticipatory synchronization in conversation, we should spend effort finding ways to model in a sense the rhythm of the speaker in order to predict a timing to which the interlocutor can synchronize. This model needs not necessarily be in terms of a rhythm or beat pulse, but can possibly also be in terms such as turn shift latencies (investigated by Nagaoka et al. [27]) or the oscillators of Wilson and Wilson [49]. The quote from Bréal in Section 2, about listeners who in a sense ‘speak along’ with the speaker, could form a metaphor for a way of modeling speaker behavior that includes anticipation as a core element.

5 Architectural Consequences

In the preceding sections we discussed some aspects of coordination apparent in human-human conversation and three of our applications in which coordination between a virtual human and the human user is a central theme. After coming to the conclusion that it is worthwhile to pursue a way of designing anticipatory synchronization for virtual humans, we now turn to our work on the development of an architecture for flexible generation of coordinated multimodal behavior. We have worked on this architecture in the context of many different applications where the expressive behavior of virtual humans needed to be specified and executed. Initially it was mostly concerned with coordination between the different modalities within the behavior expressions [47]. Over time, coordination with the behavior of a virtual human's interlocutor came to play a greater role. In this section we review our current architecture and foreseen developments with a special focus on coordination and anticipatory synchronization.

5.1 The Behavior Markup Language

Currently, our architecture for specifying and executing expressive behavior for a virtual human is based on the Behavior Markup Language (BML) [45]. In BML, multimodal expressions are composed of *behaviors*: instances of a certain 'action' by a virtual human on a single modality (for example a single pointing gesture, or a spoken sentence). Behaviors contain different *phases*. *Synchronization points* are defined at the bounds of the behavior phases. Figure 6 shows the mandatory phases and synchronization points of a BML behavior, but specific behaviors can also be defined with custom synchronization points, and thus custom phases. The duration of a phase can be 0. Coordination is achieved simply by specifying the alignment of synchronization points in gestures, speech and possibly other 'behaviors' in other modalities. A *sequence* of behaviors can be specified using 'before' and 'after' constraints. An 'after' constraint e.g. specifies that a synchronization point of one behavior should occur after a synchronization point in another behavior. The BML specification leaves it up to the modules

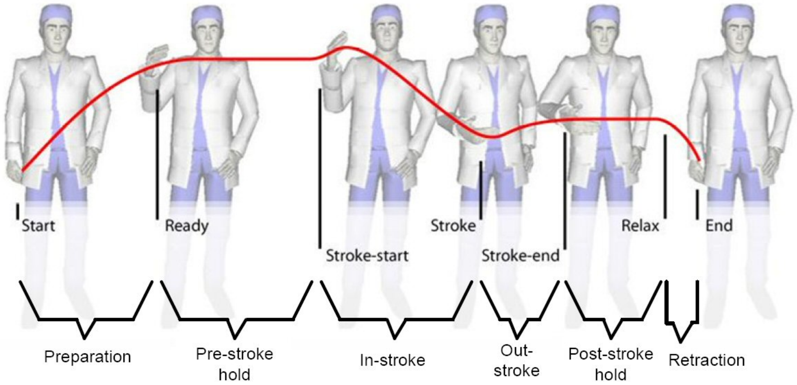


Fig. 6. Mandatory gesture phases and synchronization points in BML, picture modified from Kopp et al. [23]

that actually generate the animations (the BML realizer) to determine how long after the synchronization point that is. Other such constraints can specify that synchronization points of behaviors should occur at a certain absolute time or that two or more synchronization points of different behaviors should occur at the same time. A leading modality is not required, complex coordination is achieved at generation time (see also the inset "Multimodal Coordination in Virtual Humans: A Brief History").

5.2 External Synchronization

Synchronization can also align behavior with events occurring in the world 'outside' the virtual human. The exact form that such alignment takes depends on the type of event, and its predictability. For some events, one can 'predict' that they might occur, but not when (e.g., another person entering the room, or suddenly offering to shake your hand). Currently BML can specify synchronization to such events in a reactive way: if the designer included the appropriate reactions to somebody offering to shake hands, BML allows one to plan that reaction using an event/wait system described below. For other events, it might be possible to predict the timing with which they occur, such as for some pauses in speech. Below, we introduce the synchronization of behaviors to timing-predictable events in the outside world using the BML^{T} 'observer' extension. (As a third case, there are of course also events outside the domain modeled by the system, such as an escaped tiger entering the office. Clearly, most virtual humans will not be able to sensible coordinate their behavior with tigers entering the office).

Multimodal Coordination in Virtual Humans: A Brief History

In classic multimodal systems [9,10,42,28], speech and gesture are coordinated by timing the gestures to speech generated by a speech synthesizer. Speech then guides the timing of the gestures; speech is the leading modality [47]. However, speech/gesture timing in humans shows more complex coordination. For example: speech output can be delayed so that a complex gesture can be finished or a gesture's hold phase can be used to correct the timing of a gesture that was started too early [24].

MURML [22] is the first gesture and speech synthesis system that does not require speech as a leading modality. Multimodal behavior is planned as a concatenation of chunks, based on McNeill's [26] segmentation hypothesis. The chunks contain a segment of speech and/or one gesture, which are timed at their synchronization points. If only hand gestures and speech are to be coordinated, such an approach works fine. However, in many multimodal generation applications the behavior of virtual humans is not confined to just gesture and speech: virtual humans could operate and discus the working of complex machinery [19], behave naturally in a war zone [35], comment on an ongoing soccer match [1], or simply walk through an environment while conversing.

To support the specification of the coordination of such a wider range of modalities, we designed MultiModalSync [47]. MultiModalSync implicitly defines a leading modality, but this leading modality can change over time. Such a change does not emerge from the behavior generation itself, it has to be specified beforehand.

In BML [45], virtual human researchers (including those who designed the systems mentioned above), collaborate to provide a framework in which multimodal alignment to both internal modalities and external events can be specified in a flexible way.

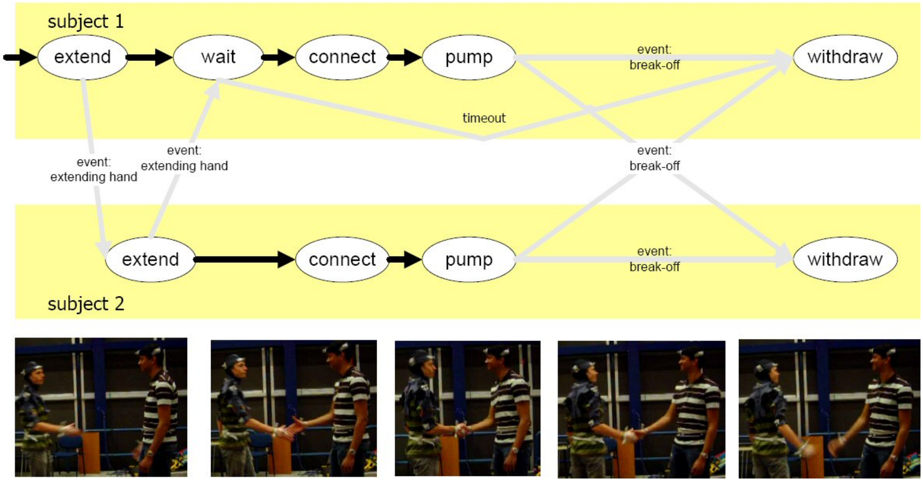
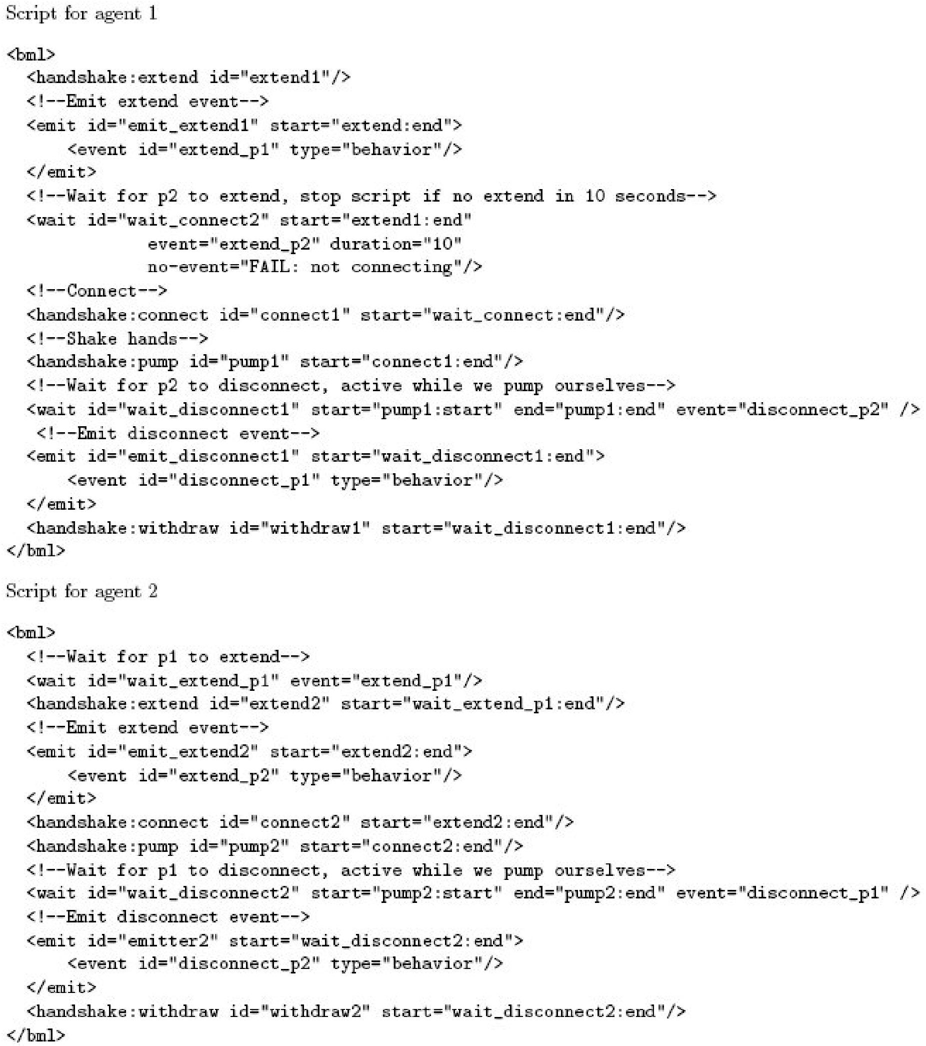


Fig. 7. Two agents that run a BML script shaking hands. Synchronization between scripts is achieved using the BML event/wait system.

Reactive coordination. Using an event/wait system, BML can be used to specify a script segment that handles a specific event that might occur during a specified time period in the scripts execution. Such events come from the outside, and possibly from other agents (executing their own BML scripts).

Figure 7 shows how the event/wait system can be used to model a handshake between two agents, independently running their own BML script. Possible scripts for the handshake are shown in Figure 8. Although these scripts form a good starting point for reactive coordination with an interlocutor, they suffer from the fact that there is no monitoring feedback loop that allows one agent to use predictions of the behavior of the other to coordinate his behavior. In the example script, one of them extends his hand, and then waits for the other to grasp that hand. Normally when two people shake hands, they will coordinate the timing of their behavior in such a way that they together arrive at the 'connect' point.

Observing and predicting. Currently we are collaborating on extending BML with possibilities to align multimodal behavior to predicted timing of external events. Predictions can concern physical events (predict where the ball will be in order to be able to catch it, or predict the beat in music) as well as social events (predict behavior of conversational partner in order to plan and synchronise contingent behavior). Such synchronization can not be achieved by the event/wait system described in core BML since BML events are, by design, non-repeatable and unpredictable. Therefore, the BML^{T} 'observer' extension is developed to provide coordination with outside world events. An 'observer' is a module that provides predictions of the timing of events in the form of synchronisation points, the exact time stamp of which can dynamically be modified with updated predictions. In the dancer for example, such synchronisation points are made available for predictions of the beat of the music. The dance movements can be specified as being aligned to those predictions (cf. Figure 4).



Script for agent 1

```
<bml>
  <handshake:extend id="extend1"/>
  <!--Emit extend event-->
  <emit id="emit_extend1" start="extend:end">
      <event id="extend_p1" type="behavior"/>
  </emit>
  <!--Wait for p2 to extend, stop script if no extend in 10 seconds-->
  <wait id="wait_connect2" start="extend1:end"
              event="extend_p2" duration="10"
              no-event="FAIL: not connecting"/>
  <!--Connect-->
  <handshake:connect id="connect1" start="wait_connect:end"/>
  <!--Shake hands-->
  <handshake:pump id="pump1" start="connect1:end"/>
  <!--Wait for p2 to disconnect, active while we pump ourselves-->
  <wait id="wait_disconnect1" start="pump1:start" end="pump1:end" event="disconnect_p2" />
   <!--Emit disconnect event-->
  <emit id="emit_disconnect1" start="wait_disconnect1:end">
      <event id="disconnect_p1" type="behavior"/>
  </emit>
  <handshake:withdraw id="withdraw1" start="wait_disconnect1:end"/>
</bml>
```

Script for agent 2

```
<bml>
  <!--Wait for p1 to extend-->
  <wait id="wait_extend_p1" event="extend_p1"/>
  <handshake:extend id="extend2" start="wait_extend_p1:end"/>
  <!--Emit extend event-->
  <emit id="emit_extend2" start="extend2:end">
      <event id="extend_p2" type="behavior"/>
  </emit>
  <handshake:connect id="connect2" start="extend2:end"/>
  <handshake:pump id="pump2" start="connect2:end"/>
  <!--Wait for p1 to disconnect, active while we pump ourselves-->
  <wait id="wait_disconnect2" start="pump2:start" end="pump2:end" event="disconnect_p1" />
  <!--Emit disconnect event-->
  <emit id="emitter2" start="wait_disconnect2:end">
      <event id="disconnect_p2" type="behavior"/>
  </emit>
  <handshake:withdraw id="withdraw2" start="wait_disconnect2:end"/>
</bml>
```

Fig. 8. The handshake BML script for agent 1 and agent 2

5.3 An Architecture That Supports Anticipatory Coordinated Multimodal Interaction

Figure 9 summarizes the overall architecture of our system. The scheduler is responsible for the alignment of the behaviors on different modalities. It generates an execution plan, based on the provided BML script. The execution of a BML^{T} script requires dynamic (re)-planning capabilities, since the predicted times are not fully known beforehand and can and will be updated during the execution of the script. For example, in the dancer the timing of dance moves is adapted so that it is aligned to changing predictions of the beat of the music, while the dance animation is being executed. Replanning can involve

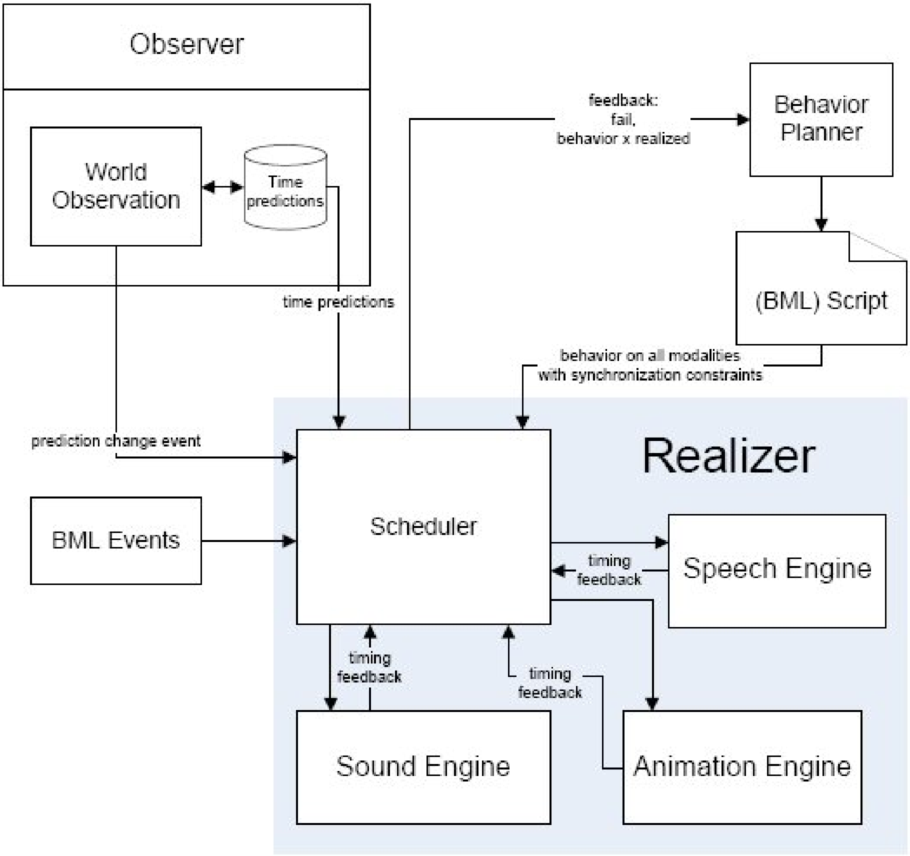


Fig. 9. Architecture

1. translating the start time of behaviors
2. stretching or skewing behaviors to fit timing constraints
3. skip behaviors that have low priority (as indicated by the BML script)
4. or, if the methods above fail, inform the behavior planner that generated the BML script, which in turn can provide an alternative script

Figure 10 shows some of the strategies that can be used to retime behaviors so that timing constraints are met. Many more retiming strategies are possible. Retiming by stretching and/or skewing behaviors raises two issues:

1. How can a behavior be stretched/skewed?
2. Which of the behaviors have to be stretched/skewed by how much?

Different modalities stretch and skew in different ways. Our previous research explored the stretching and skewing of exercise motion [48]. While a timing constraint specifies the timing of only one behavior phase, a stretch/skew operation possibly influences the timing of all phases in the behavior.

The alignment of behaviors is specified using time constraints. To satisfy these constraints, one or more of the behaviors that are to be aligned have to be stretched or

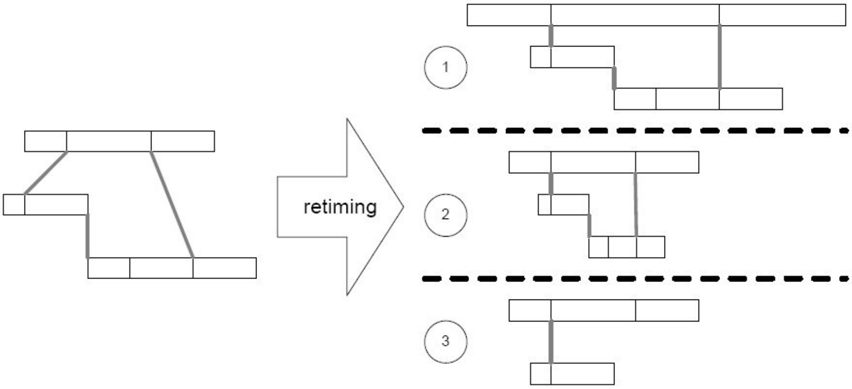


Fig. 10. Retiming is necessary to align the three behaviors (indicated by the 3 rectangles), executed at different modalities. Retiming strategy 1 stretches the first behavior. Strategy 2 skews the second and third behaviors. Strategy 3 omits the third behavior.

skewed by a certain amount. Typically, an infinitely large set of different stretch/skew amounts on the behaviors that are to be aligned can satisfy the alignment time constraints. The scheduler should select one of these. If a cost can be calculated for each stretch/skew amount on each behavior, then the scheduler can select stretch/skew amounts for each behavior in such a way that the total cost is minimized.

Finally, each output modality has its own low-level planner. Given timing constraints on a behavior, the previous behavior(s) on this modality and the behavior to be executed, the planners can provide timing information for each behavior phase, the cost of the behavior and whether or not the behavior can be executed satisfying the provided constraints. The scheduler can query the planners for this information, attempting different behavior timing combinations, until an execution plan is found that has minimal cost.

6 Conclusions

In this paper we reported about our research in progress on mutually coordinated anticipatory multimodal interaction. We looked at three applications of this research: an interactive virtual dancer, an interactive virtual conductor, and an interactive virtual trainer. In these applications expressive behavior of a virtual human has to be synchronized with external events and predictions that are related to the performance of the virtual human or its perception of the environment. We discussed anticipation in conversations and continuous synchronizing expressive behavior in conversations based on expectations. Specification of synchronization is investigated with the Behavioral Markup Language (BML) and extensions of this language that are introduced to accommodate synchronization required for our applications. Architectural consequences for a virtual human that has to act human-like have been discussed. In particular the issue of replanning and retiming of actions to be performed by a virtual human need to be addressed.

In our future research we will investigate more examples of natural continuous multimodal interaction that appear in real-life situations. That is, interactions in 'ambient intelligence' environments, where the environments or their virtual inhabitants have to deal with human partners in charge of multiple related and unrelated tasks (e.g., in a smart home office environment) or where the virtual humans have to take into account that their human partners are also involved in family related activities (e.g., in a smart home environment) that have impact on the planning and timing of their expressive behavior in their interaction with virtual partners.

Acknowledgements. This work has been supported by funding from the European IST Programme Project FP6-033812 (Augmented Multi-party Interaction, publication AMIDA-82), the COST Action 2102 (Cross-Modal Analysis of Verbal and Non-verbal Communication), and the GATE project, funded by the Netherlands Organization for Scientific Research (NWO) and the Netherlands ICT Research and Innovation Authority (ICT Regie). This paper only reflects the authors' views and funding agencies are not liable for any use that may be made of the information contained herein.

References

1. André, E., Rist, T., van Mulken, S., Klesen, M., Baldes, S.: The automated design of believable dialogues for animated presentation teams. In: Cassell, J., Prevost, S., Sullivan, J., Churchill, E. (eds.) Embodied Conversational Agents, pp. 220–255. MIT Press, Cambridge (2000)
2. Bailenson, J.N., Yee, N.: Digital chameleons: Automatic assimilation of nonverbal gestures in immersive virtual environments. Psychological Science 16(1), 814–819 (2005)
3. Basu, S.: Conversational scene analysis. MIT Press, Cambridge (2002)
4. Bavelas, J.B., Coates, L., Johnson, T.: Listeners as co-narrators. Journal of Personality and Social Psychology 79(6), 941–952 (2000)
5. Boker, S.M., Xu, M., Rotondo, J.L., King, K.: Windowed cross-correlation and peak picking for the analysis of variability in the association between behavioral time series. Psychological Methods 7(3), 338–355 (2002)
6. Bos, P., Reidsma, D., Ruttkay, Z.M., Nijholt, A.: Interacting with a virtual conductor. In: [16], pp. 25–30
7. Bull, M.: An analysis of between-speaker intervals. In: Proceedings 1996 of the Edinburgh Postgraduate Conference in Linguistics and Applied Linguistics, pp. 18–27 (1996)
8. Carletta, J.C., Ashby, S., Bourban, S., Flynn, M., Guillemot, M., Hain, T., Kadlec, J., Karaiskos, V., Kraaij, W., Kronenthal, M., Lathoud, M., Lincoln, M., Lisowska, A., McCowan, I., Post, W.M., Reidsma, D., Wellner, P.: The AMI meeting corpus: A preannouncement. In: Renals, S., Bengio, S. (eds.) MLMI 2005. LNCS, vol. 3869, pp. 28–39. Springer, Heidelberg (2006)
9. Cassell, J., Pelachaud, C., Badler, N., Steedman, M., Achorn, B., Becket, T., Douville, B., Prevost, S., Stone, M.: Animated conversation: rule-based generation of facial expression, gesture & spoken intonation for multiple conversational agents. In: SIGGRAPH 1994: Proceedings of the 21st annual conference on Computer Graphics and Interactive Techniques, pp. 413–420. ACM Press, New York (1994)
10. Cassell, J., Vilhjálmsson, H.H., Bickmore, T.: BEAT: The behavior expression animation toolkit. In: Fiume, E. (ed.) SIGGRAPH 2001, Computer Graphics Proceedings, pp. 477–486. ACM Press, New York (2001)
11. Coates, J.: No gap, lots of overlap: turn-taking patterns in the talk of women friends. Multilingual Matters, 177–192 (1994)
12. Cowley, S.J.: Of timing, turn-taking, and conversations. Journal of Psycholinguistic Research 27(5), 541–571 (1998)

13. Crown, C.L.: Coordinated Interpersonal Timing of Vision and Voice as a Function of interpersonal Attraction. Journal of Language and Social Psychology 10(1), 29–46 (1991)
14. Emmott, S.J., Travis, D.: Information superhighways: multimedia users and futures. Academic Press, Inc., Duluth (2005)
15. Goodrich, S., Henderson, L., Allchin, N., Jeevaratnam, A.: On the peculiarity of simple reaction time. The Quarterly Journal of Experimental Psychology Section A 42(4), 763–775 (1990)
16. Harper, R., Rauterberg, M., Combetto, M. (eds.): 5th International Conference on Entertainment Computing. LNCS, vol. 4161. Springer, Heidelberg (2006)
17. Heylen, D., Nijholt, A., Poel, M.: Generating nonverbal signals for a sensitive artificial listener. In: Esposito, A., Faundez-Zanuy, M., Keller, E., Marinaro, M. (eds.) COST Action 2102. LNCS (LNAI), vol. 4775, pp. 264–274. Springer, Heidelberg (2007)
18. Izdebski, K., Shipp, T.: Minimal reaction times for phonatory initiation. Journal of Speech and Hearing Research 21(4), 638–651 (1978)
19. Johnson, L.L., Rickel, J.W., Lester, J.: Animated pedagogical agents: Face-to-face interaction in interactive learning environments. International Journal of Artificial Intelligence in Education 11, 47–78 (2000)
20. Jonsdottir, G.R., Gratch, J., Fast, E., Thórisson, K.R.: Fluid semantic back-channel feedback in dialogue: Challenges and progress. In: [27], pp. 154–160
21. Keller, E.: Beats for individual timing variation. In: Esposito, A., Keller, E., Marinaro, M., Bratanic, M. (eds.) The Fundamentals of Verbal and Non-verbal Communication and the Biometrical Issue. NATO Security through Science: Human and Societal Dynamics, vol. 18, pp. 115–128. IOS Press, Amsterdam (2007)
22. Kopp, S.: Surface realization of multimodal output from xml representations in MURML. In: Invited Workshop on Representations for Multimodal Generation (2005)
23. Kopp, S., Krenn, B., Marsella, S., Marshall, A.N., Pelachaud, C., Pirker, H., Thórisson, K.R., Vilhjálmsson, H.H.: Towards a common framework for multimodal generation: The behavior markup language. In: Gratch, J., Young, M.R., Aylett, R., Ballin, D., Olivier, P. (eds.) IVA 2006. LNCS (LNAI), vol. 4133, pp. 205–217. Springer, Heidelberg (2006)
24. Kopp, S., Wachsmuth, I.: Model-based animation of co-verbal gesture. In: CA 2002: Proceedings of the Computer Animation Conference, p. 252. IEEE Computer Society, Washington (2002)
25. Maatman, R.M., Gratch, J., Marsella, S.: Natural behavior of a listening agent. In: Panayiotopoulos, T., Gratch, J., Aylett, R., Ballin, D., Olivier, P., Rist, T. (eds.) Intelligent Virtual Agents. Lecture Notes in Computer Science, vol. 3661, pp. 25–36. Springer, Berlin (2005)
26. McNeill, D.: Hand and Mind: What Gestures Reveal about Thought. University of Chicago Press, Chicago (1995)
27. Nagaoka, C., Komori, M., Yoshikawa, S.: Synchrony tendency: interactional synchrony and congruence of nonverbal behavior in social interaction. In: Proceedings International Conference on Active Media Technology, pp. 529–534 (2005)
28. Noot, H., Ruttkay, Z.: The Gestyle language. In: International workshop on gesture and sign language based human-computer interaction (2003)
29. O'Connell, D.C., Kowal, S., Kaltenbacher, E.: Turn-taking: A critical analysis of the research tradition. Journal of Psycholinguistic Research 19(6), 345–373 (1990)
30. Pelachaud, C., Martin, J.-C., André, E., Chollet, G., Karpouzis, K., Pelé, D. (eds.): Intelligent Virtual Agents, 7th International Conference. LNCS, vol. 4722. Springer, Heidelberg (2007)
31. Ramseyer, F., Tschacher, W.: Synchrony: A Core Concept for a Constructivist Approach to Psychotherapy. Constructivism in the Human Sciences 11(1), 150–171 (2006)

32. Ramseyer, F., Tschacher, W.: Synchrony in dyadic psychotherapy sessions. In: Simultaneity: Temporal Structures and Observer Perspectives, ch. 18. World Scientific, Singapore (to appear, 2008)
33. Reeves, B., Nass, C.: The media equation: how people treat computers, television, and new media like real people and places. Cambridge University Press, New York (1996)
34. Reidsma, D., Welbergen, H., van Poppe, R., Bos, P., Nijholt, A.: Towards bidirectional dancing interaction. In: [16], pp. 1–12
35. Rickel, J.W., Gratch, J., Marsella, S., Swartout, W.: Steve goes to Bosnia: Towards a new generation of virtual humans for interactive experiences. In: AAAI Spring Symposium of Artificial Intelligence and Interactive Entertainment (2001)
36. Robins, B., Dautenhahn, K., Nehaniv, C.L., Mirza, N.A., Francois, D., Olsson, L.: Sustaining interaction dynamics and engagement in dyadic child-robot interaction kinesics: Lessons learnt from an exploratory study. In: Proc. of the 14th IEEE International Workshop on Robot and Human Interactive Communication, RO-MAN 2005 (2005)
37. Ruttkay, Z.M., Zwiers, J., Welbergen, H., van Reidsma, D.: Towards a reactive virtual trainer. In: Gratch, J., Young, M., Aylett, R.S., Ballin, D., Olivier, P. (eds.) IVA 2006. LNCS (LNAI), vol. 4133, pp. 292–303. Springer, Heidelberg (2006)
38. Sacks, H., Schegloff, E.A., Jefferson, G.: A simplest systematics for the organization of turn-taking for conversation. Language 50(4), 696–735 (1974)
39. Sanders, C.: The Paris years. In: Sanders, C. (ed.) The Cambridge Companion to Saussure, Ch. 2., pp. 30–46. Cambridge University Press, Cambridge (2005)
40. Slowiaczek, L.M.: Semantic priming in a single-word shadowing task. The American Journal of Psychology 107(2), 245–260 (1994)
41. Suzuki, N., Takeuchi, Y., Ishii, K., Okada, M.: Effects of echoic mimicry using hummed sounds on human-computer interaction. Speech Communication 40(4), 559–573 (2003)
42. Theune, M., Heylen, D., Nijholt, A.: Generating Embodied Information Presentations. In: Stock, O., Zancanaro, M. (eds.) Multimodal Intelligent Information Presentation, Ch. 3. Kluwer Series on Text, Speech and Language Technology, vol. 27, pp. 47–70. Kluwer Academic Publishers, Dordrecht (2005)
43. Thórisson, K.R.: Communicative humanoids: a computational model of psychosocial dialogue skills. PhD thesis, MIT Media Laboratory (1996)
44. Thórisson, K.R.: Natural Turn-Taking Needs No Manual: Computational Theory and Model, from Perception to Action. In: Multimodality in Language and Speech Systems, pp. 173–207. Kluwer Academic Publishers, Dordrecht (2002)
45. Vilhjálmsson, H.H., Cantelmo, N., Cassell, J., Chafai, N.E., Kipp, M., Kopp, S., Mancini, M., Marsella, S., Marshall, A.N., Pelachaud, C., Ruttkay, Z.M., Thórisson, K.R., van Welbergen, H., van der Werf, R.J.: The behavior markup language: Recent developments and challenges. In: [30], pp. 99–111
46. Ward, N., Tsukahara, W.: A Responsive Dialog System. In: Wilks, Y. (ed.) Machine Conversations, pp. 169–174. Kluwer Academic Publishers, Dordrecht (1999)
47. Welbergen, H., van, N.A., Reidsma, D., Zwiers, J.: Presenting in virtual worlds: Towards an architecture for a 3D presenter explaining 2D-presented information. IEEE Intelligent Systems 21(5), 47–53 (2006)
48. Welbergen, H., van Ruttkay, Z.: On the parameterization of clapping. In: Proc. 7th International Workshop on Gesture in Human-Computer Interaction and Simulation (to appear, 2007)
49. Wilson, M., Wilson, T.P.: An oscillator model of the timing of turn-taking. Psychonomic Bulletin & Review 12(6), 957–968 (2005)
50. Yngve, V.H.: On getting a word in edgewise. In: Papers from the 6th Regional Meeting of the Chicago Linguistics Society, pp. 567–577. University of Chicago (1970)