Computational models of social and emotional turn-taking for embodied conversational agents: a review

Rieks op den Akker and Merijn Bruijnes

University of Twente, 7500AE Enschede, the Netherlands

Abstract

The emotional involvement of participants in a conversation not only shows in the words they speak and in the way they speak and gesture but also in their turn-taking behavior. This paper reviews research into computational models of embodied conversational agents. We focus on models for turn-taking management and (social) emotions. We are particularly interested in how in these models emotions of the agent itself and those of the others influence the agent's turn-taking behavior and vice versa how turn-taking behavior of the partner is perceived by the agent itself. The system of turn-taking rules presented by Sacks, Schegloff and Jefferson (1974) is often a starting point for computational turn-taking models of conversational agents. But emotions have their own rules besides the "one-at-a-time" paradigm of the SSJ system. It turns out that almost without exception computational models of turn-taking behavior that allow "continuous interaction" and "natural turn-taking" do not model the underlying psychological, affective, attentional and cognitive processes. They are restricted to rules in terms of a number of superficially observable cues. On the other hand computational models for virtual humans that are based on a functional theory of social emotion do not contain explicit rules on how social emotions affect turn-taking behavior or how the emotional state of the agent is affected by turn-taking behavior of its interlocutors. We conclude with some preliminary ideas on what an architecture for emotional turn-taking should look like and we discuss the challenges in building believable emotional turn-taking agents.

A COMMIT deliverable *May 23, 2012*

Contents

1. Introduction

Life is emotion. Emotions make us move, fight or flight, approach or withdraw. Some emotions, such as fear, anger, and disgust refer to the basic values of our lives. These emotions sometimes take over the control of our behavior, in particular when the situation asks an immediate response. Sometimes for good, sometimes for bad. People can recognize feelings that others have, because they show in the human face, or in the way someone speaks, or behaves. But also seeing an image of a human face, we recognize happiness or sadness. People play emotions in theater, in movies, but also in their daily life. In cartoons and movies we can recognize characters having certain emotions.

Having conversations makes up a good deal of our social lives. The emotions of people having a conversation not only show in the content of their contributions, in their gestures, body postures and expressions but also in the moments they talk, or do not talk, in their "turn-taking" behavior. In many cultures people in conversation try to follow the convention of "one-talks-at-a-time" but when the temperature of the conversation raises they often do not adhere anymore to this rule. In an animated conversation people talk at the same time. When someone is attacking someone verbally we can imagine that the addressed one becomes angry if he feels that he is treated unfair. It is believable for us that the anger makes the insulted addressee defend himself by trying to stop the speaker talking. Sometimes people apologize for interrupting the speaker. Apparently because they feel they did behave in an impolite way. What makes people having a conversation act the way they do? In what situations do emotions (more than rational decisions) make listeners interrupt the speaker, and say what they say the way they say it? And when, after the fact, do they feel sorry for it? And, when do they apologize for their behavior?

In many professions a great part of the work consists of having conversations. Think of the policeman who has to interrogate a suspect, the street worker who has to negotiate with a youth group, the practitioner who has to deliver the outcome of the medical search to his patient. They apply "conversational techniques" (Dutch: gesprekstechnieken) and develop conversational skills (Dutch: gespreksvaardigheden) to increase the chance that their conversations have the desired outcome. This outcome is not only that they give or receive some information, or that they motivate or warn their clients to behave in a certain way, more importantly they want to build on qualities of a social relation that brings them further. Rules and procedures how to handle, what stance to take ("be an active listener", "show empathy"), or how empathy is expressed (say 'we', not 'you'!) are trained by means of role play.[1] Can virtual humans play a role in training scenarios with similar learning effects as human players? This requires that virtual humans understand what goes on in the minds of their conversational partners. They should also be able to express and understand the emotions that play in a given situation. They have to be able to predict the effect of their own stance on the stance that the other takes.[2] One of the basic conversational skills is to know when to take the turn and what it means when someone starts talking out of turn, or remains silent when he is expected to give an answer.

Technology is based on our ability to reflect, detach, objectify and combine aspects from the live stream of events. Language technology, for example, is possible because of the abstraction of the language from the actual and meaningful expression of thought by some speaker in a practical encounter with other humans. We can observe what people say or do "from the outside". We can detach from the actual event as it is experienced the abstract form of a sequence of sounds, and record it. So we can transcend "what happens" as something in itself from the stream of life. This allows us to transpose the imaginations of our intellect to other times and reconstruct, reproduce them. We can thus say something without meaning it, or let others say something without meaning it. Our intellect is able to distinguish such things as the "truth value", the "propositional content", "the sound", "the grammatical form", and the organ that produces the sound. In the same way we can abstract and objectify such things as "small behaviors" from the concrete person that we meet in a social encounter.[3] We can classify and la-

[1] For a Dutch group that organizes training sessions for learning how to deal in "complex conversations" see www.wildekastanje.nl

[2] See for example Timothy Leary's Rose on the effect of our acting on social relationships.

[3] Goffman defines the subject of interaction analysis as those "small behaviors" such as an eye blink, a head movement, etc. that people make in conversations. "The subject matter however, can be identified. It is that class of events which occur during co-presence and by virtue of co-presence. The ultimate behavioral materials are the glances, gestures, positionings, and verbal statements that people continuously feed into the situation whether intended or not. These are the external signs of orientation and involvement - states of mind and body not ordinarily examined with respect to their social organization." (Introduction of Goffman (1967)).

bel these behaviors as we do with flowers or insects. A person then becomes "someone who shows these behaviors". We can copy these behaviors as we can copy images and sounds. We can reproduce the sentences and let them be pronounced by a speaker, a device, making the impression of a truthful and original expression, a statement that reflects a position that someone takes. This impression of a genuine encounter with a real person can be very strong even though we know that these fabricated speakers are only virtual. This impression works in the same way as the image remind us of the original event. The photo of our relatives makes them present for us, the moment when the photo is presented to us, because of some memorized resemblance. The voice of the answering machine not only presents the person that we recognize but also pretends the actual presence of the person whose voice we recognize. In this technological world of make belief and experience design we work on the construction of believable animated conversational agents, graphical as well as robots, that "pretend" to have emotions and that try to become inhabitants of our social life.

The idea of a "believable character" originates from theater and other media arts. "It does not mean an honest or reliable character but one that provides the illusion of life and thus permits the audiences suspension of disbelief."(Bates 1994). "Artificial intelligence researchers trying to create engaging apparently living creatures may find important insight in the work of artists who have explored the idea of believable characters. In particular.. appropriately timed and clearly expressed emotion is a central requirement for believable characters." Bates (1994) quotes from a classic work on Disney animations.

> Disney animation makes audiences really believe in characters whose adventures and misfortunes make people laugh and even cry. There is a special ingredient in our type of animation that produces drawings that appear to think and make decisions and act of their own volition it is what creates the illusion of life.

The first and most important lesson learned from the artists is the importance of emotion expressions in these characters. "The apparent desires of a character and the way the character feels about what happens in the world with respect to those desires are what make us care about that character. If the character does not react emotionally to events if they don't care then neither will we. The emotionless character is lifeless as a machine." Goals and needs and the agent's appraisals of events with respect to these goals and needs are key to producing a clearly defined emotional state in the creature. The situation in which the creature is affectively involved should be made clear to the spectator.

Virtual humans are computer animations of human characters that can play roles in specific scenarios. They interact and have conversations with other virtual humans or with real humans. These animated embodied conversational agents (ECAs) can have a realistic or cartoon-like body and they can engage in spoken discourse and dialogue. They can use voice with appropriate prosody and intonation, synchronize their mouth movements to the words uttered, make gestures, assume postures, and produce facial expression and communicative gaze behavior Poggi et al. (2005). These animated conversational characters are increasingly used in a wide range of different application areas, including virtual training environments (Traum et al. (2005), Prendinger and Ishizuka (2001)), in task-oriented dialog systems representing agents that help users find their way in virtual environments (Hofs et al. (2010)), in storytelling systems (Theune et al. (2004)),in serious games such as Siren, a multi-player game for learning how to resolve social conflicts Yannakakis et al. (2010), as well as in e-commerce applications where computer agents play a role as sales assistant (Gebhard et al. (2003)).

Some authors see "special links, or bonds" between users and ECAs, in applications where the ECA functions as the presentation of a personal coach or companion[4] (Bickmore and Picard (2005) and Shearer et al. (2007)). For establishing such a long term relationship engagement, cognitive and emotional involvement as well as commitment are key factors. If this is the case, then for an ECA to be able to establish and maintain relations, it must be endowed with mechanisms that allow it to perceive, adapt to and generate behaviors relating to attention and emotional involvement (Peters et al. (2005)).

Artificial embodied conversational agents should show appropriate and coherent emotional and emphatic behavior and know the social rules of turn-taking. They should be able to regulate the flow of the conversation showing

[4] "By Companions we mean conversationalists or confidants not robots but rather computer software agents whose function will be to get to know their owners over a long period. Those may well be elderly or lonely, and the contributions in the book focus not only on assistance via the internet (contacts, travel, doctors etc.) but also on providing company and Companionship, by offering aspects of real personalization." (Wilks (2009))

the gestures and expressions that fit their emotional state and the situation. Just like humans such agents are able to do many things "at the same time". When speaking they can observe what listeners do and react on the feedback they give. They can stop speaking immediately when some event suddenly requires immediate attention. These computer animations require computer architectures based on complex models of agents that show emotional conversational behavior when interacting with other virtual or real characters.

In this report we discuss computational models of turn taking and emotions in conversations as well as architectures and systems that are based on these models. Since the computational turn in psychology it is a well-established idea that we "understand" how the human mind "works" by building computational models and by implementing these models as computer programs. We do not discuss here the principal psychological and philosophical issues raised by the view that the mind is a computational system, what Margareth Boden called the computational metaphor (Boden (1979)). We are interested in these models because we want to make computer agents that have build in conversational skills. These animated conversational characters show "believable emotional and conversational behavior" in the eyes of humans that observe or interact with these agents. We build systems that make believable impressions so that humans get engaged in a dialogue with these agents. We do this in such a way that humans react verbally and non-verbally in a social way to the acts of these agents. The criterium for success of our work is how these animations are perceived by humans in some situation and scenario of use, measured by how they assess and response to these artifacts' behaviors. This type of animations may show their practical value beyond their value to entertain, in the same way as images, statues, video recordings, theater and role plays have shown their value in everyday life.[5]

If we say that virtual humans "have emotions", "talk" or "behave" we

[5] "What minimal model of the actor is needed if we are to wind him up, stick him in amongst his fellows, and have an orderly traffic of behavior emerge?" This formulation used by Goffmann in the Introduction of his Interaction Rituals (1967!) to present the main theme of the bundle of essays already shows that for the researcher who studies these abstract "small behaviors" the actor could as well be a real human as an artificial virtual agent. It's primarily "the moments" that count, not the men. The question is how this abstraction works out if we apply these artificial social agents. A question we will not consider here.

mean that in a metaphorical sense, not in the genuine sense of the word. A stone doesn't know Newton's law, a plant doesn't know it's spring time, and puppets don't cry. It is normal language use to say that computers or systems act or do something, or even behave in a certain way, even if we know that we don't mean that literally as if this implies that they are the active subjects of these activities in the way humans are the authors and actors of their genuine behaviors. Picard and Klein put it this way:

> When we refer to "affect perception" or to "recognition of affect" we do not intend to imply that machines are conscious or human-like in how they perform such tasks. Our usage is one of convenience; we lack a better short phrase to replace what we really mean by machine affect recognition: "a computer system employing techniques such as signal sensing and detection, pattern analysis, probabilistic inference, and dynamic reasoning, in order to extract and characterise relevant patterns of sensed data in a way that produces a result similar to what a human would have produced if he or she had tried to observe and characterise the inputs according to their affective qualities." (Picard and Klein (2002), p.4)

It is difficult to avoid this analogical and metaphorical language use when talking about technology. The reason for this is that this technology works precisely by virtue of the fact that what the physical processes stand for are the meaningful signals and patterns that makes their very sense for us as designers and users of this technology. Moreover, that we don't see our own activities as the causes of the work of the machine, but instead conceive this working as something that is done "by the machine itself" is because of the abstract -representational and arbitrary- relation between our acts (the inputs to the system) and what the machine does and what it produces as output. The computer is based on a relation between our acts and the computation which is only representational: what we do when we set the machine to calculate cannot be understand as the cause of the computational work done by the machine itself.

Ethical issues related to misconceptions and confusions caused by these artifacts or by abuse of this type of artifacts need to be considered but are not at stake here.

1.1. Organisation of this report

Turn can be intuitively defined as "the talk of one party bounded by the talk of others". But in "it's my turn now" turn refers to a social convention. Moreover, can non-verbal acts also fill in a turn space? Or silence for that matter. The notion of turn, which is central in turn-taking theory, is as basic as it is debated. In section 2 we discuss the turn taking model and we review the critical comments that it received from various research areas.

In section 3 we discuss two architectures for turn-talking. One is based on the turn taking model, the second alternative is based on the idea of "continuous multi-modal interaction" (Reidsma et al. (2010)) and the fact that the agent while speaking should also "simultaneously" be attentive to the addressees, who continuously provide him with feedback on the words he is saying. See Simon (1967) for an early account of what this implies for AI and Akker and Heylen (2007) for the analysis of "feedback loops" in conversations.

In section 4 we discuss what it means for virtual humans to have emotions and why embodied conversational agents should have emotions. In section 5 we discuss emotion models. A computational model of emotion must explain both "the rapid dynamics of some emotional reactions" as well as "the slower responses that follow deliberation" (Marsella and Gratch (2009)). We discuss the relation between (social) emotions and turn taking. In section 6 we discuss architectures for emotional agents. In the section 7 we sketch a design for an architecture that incorporates the relations between emotions and turn-taking behavior. In section 8 we end with a conclusion and a discussion about the challenges in future work in building emotional natural turn taking agents.

2. The turn taking model of interaction

In this section we review the turn-taking model[6] Goodwin said:

> In the abstract, the phenomenon of turn-taking seems quite easy to define. The talk of one party bounded by the talk of others

[6] According to Duncan (1972a) the term "turn taking" has been independently suggested by Yngve, 1970, and by Goffman (in a personal communication with Duncan, June 5, 1970). Schegloff (1968) proposed the "basic rule for conversations: one party at a time", p.1076.

> constitutes a turn, with turn-taking being the process through which the party doing the talk of the moment is changed. (citation from Thorisson (2002), p.175).

However, as soon as we start to analyse natural conversational behavior, watch video recordings as Yngve and his collaborators did, the phenomena force us to reconsider this definition. We start with Yngve's idea to model the "state of mind" of the participants in a conversation and how this is related to the conversational flow (Yngve (1970)). Yngve's work and that of his followers lays ground for the kind of research needed to build artificial agents that can show natural conversational behavior in a principled way. Then we review the turn taking model of Sacks, Schegloff and Jefferson, the SSJ model (Sacks et al. (1974)). The notion of turn and the SSJ model are highly debated. In subsection 2.3 we review the main comments on the SSJ model.

2.1. Yngve: a theory of mind for turn taking

At the occasion of the sixth regional meeting of the Chicago Linguistic Society held in April 1970, Victor H. Yngve addresses the audience with a report of preliminary descriptive work within what he sees as "a new linguistic framework related to state of mind." [7]

> State of mind is postulated to contain all of the relevant contextual information, linguistic and non-linguistic, that the language user needs when carrying on communicative activity. (..) Within this broader linguistic framework, the basic research task that one faces is to discover and describe the structure of state of mind and to relate it to communicative behavior."Yngve (1970), p.567.

According to Yngve, the passing of the turn from one party to another is "nearly the most obvious aspect of conversation".(p.568). This is based on intuition as well as on the observation that the one who has the turn is more engaged in speaking activities and the part who does not have the turn is more engaged in listening activities. Be it that also those who do not have

[7] Prof. Victor Yngve is a key witness of the earliest days of computational linguistics. He was the first chairman of the Association of Computational Linguistics, founded in June 1962. Yngve was also the author of COMIT, the first string processing language.

the turn do sometimes talk, and having the turn doesn't necessarily imply actually speaking. Then, a remark is made that contains an interesting hint for an agent model.

> One might be tempted to set up a concept of turn as an aspect of or belonging to *the conversation itself as a social phenomenon.* However, we do have to treat the case where the parties to the conversation differ as to whom they *think* has the turn. In such cases it may be difficult to determine who "really" (quotes by Yngve) has the turn, if there is any such thing. Even if it could be determined who really has the turn when the parties differed in this respect, it would seem to be irrelevant for our purposes, for each person in a conversation acts according to his own concept of who has the turn, that is, according to his own state of mind, and this is all that we need in accounting for his behavior. Thus, we are led to set up, as part of the state of mind of each person in a conversation, a turn variable which takes various values, depending on who the person thinks has the turn. We this account for the difference in a person's behavior between when he thinks he has the turn and when he thinks he doesn't.

(italics by the authors)

These considerations touch the core problem of communication, since what holds for the turn holds for all variables that make up the state of mind of the partners. Yngve and his collegues observed dialogs and dialog agents. They looked at turn taking behavior in particular. What did they find?

- They confirmed that speaking and listening activities go on simultaneously and that it is common for messages to flow simultaneously in both directions between partners. Yngve's paper is the source for the notion of *the back channel*, a short reassuring message ("yes" and "uh-huh") that the one who has the turn receives without relinquishing the turn.[8]

[8] Backchannels are a sign of attentive listening. Teaching materials for learning conversational skills for professionals recommend the learner to use these to signal attentiveness and to stimulate the speaker to tell his story." From the Dutch report "JGZ-richtlijn secundaire preventie kindermishandeling": "Door het stellen van open vragen krijgen de

- They found that there is not only a difference between the roles of speaker and hearer, and the roles of having the turn or not, but there is a further distinction of *having the floor or not*. Yngve uses the term floor in the everyday sense of the word, without defining it. The meaning of the notion of floor has shift from turn to a more encompassing idea of attentive cognitive space. The notion of "floor" in conversation structure was first introduced by Sacks (1972), who regarded floor and turn at speaking as equivalent concepts (Hayashi (1991)). Edelsky (1981) studies data from faculty meetings. She was one of the first who claimed that floor and turn are *not* equivalent and defined floor as a psychologically developed, interactional space among the interactants. Hayashi (1991) revises the theory of floor that she presented previously and claims that floor is a form of community competence. "That is, it is a kind of competence that is developed in the cognitive space naturally or by mutual efforts when more than two persons interact with each other."

  About the social emotional aspects of floor Hayashi says:

  > Floor reflects social considerations of power, solidarity, cooperation, conflict, competition and the like. (...) Speaker and hearer's empathic involvement in their interlocutors and the on-going topic is one of the determinants of floor structure and management.

  (Hayashi (1991), p. 7)

- They observed that there mostly is a smooth flow of conversation, "they proceed without a hitch", with each party switching his turn variable (sic!) at the appropriate time and in the appropriate way." (The question can be asked here, how Yngve knows this, what then is the "appropriate time and place"? We think there is no other "evidence" than the "smoothness of the conversation".)

- They postulate that there are *conventional signals* that segment the speech into "paragraphs" and that are related to turn switches. "The

ouders of de jeugdige de gelegenheid om hun eigen verhaal te vertellen. Hierbij luistert de JGZ medewerker actief en stimuleert door houding, knikken, hummen et cetera." (RIVM Rapport 295001012/2010 M.M. Wagenaar-Fischer et al., 2010) only way to account for the smooth flow of conversation and the rapid and frequent switching of turns is *to suppose that there are conventional signals exchanged during conversation that function to switch the turn variables properly during conversation.*" Yngve takes the existence of these signals for granted. But, conventional signals don't come out of the blue. Where do they come from? And, how do the partners in conversation and we as outside observers know that they use these signals in the conventional way?

- An important observation that Yngve makes is that the relation between signals and meaning is not simple one to one. There is a great variability in occurrence.

- There is a link between the turn variable of the state of mind and other parts of the state of mind, such as the subject or topic of the conversation, and the listener's prior state of knowledge concerning the referent of a referential expression.

- The smooth operation of the turn change is closely associated with the obvious structural coherence of what is happening.

- The way someone interrupts or indicates he wants to say something is by means of subtle signals, such as a slight opening of the mouth and intake of breath accompanied by a slight tilting of the head. However, on some occasions it seems to work, on some occasions not.

- Concerning politeness Yngve reports that one gets the impression that there are only certain points where a polite interruption can come. *"One gets the impression of the closure of activities at various levels and that only interruptions appropriate to the level of closure would be tolerated."*

Yngve cites Erving Goffman's from his essay on Face Work:

> The conventions regarding the structure of occasions of talk represent an effective solution to the problem of organizing a flow of spoken messages. In attempting to discover how it is that these conventions are maintained in force as guides to action, one finds evidence to suggest a functional relationship between the structure of the self and the structure of spoken interaction. (Erving

> Goffman: On face work: an analysis of ritual elements in social interaction." cited from Yngve (1970), p.571)

And that is exactly what Yngve says to be interested in: "the functional relationship between the structure of the self and the structure of spoken interaction". What remains to be done is "to produce a structural description of the state of mind" (based on the structure of self of Goffman) and "show explicitly how it is related to communicative behavior in its totality, including the linguistic details." (Yngve, p.571)

Since Yngve's sketch of a new linguistic framework in 1970, a huge pile of reports on research in turn taking and floor have seen the light. Duncan (1972a) claims that there is a regular communication mechanism in our culture for managing the taking of speaking turns in face-to-face interaction. Through this mechanism, participants in an interaction can effect the smooth and appropriate exchange of speaking turns.

2.2. The SSJ model of turn taking

A frequently cited paper on turn-taking is the SSJ paper Sacks et al. (1974) in which the authors present their "simplest systematics for turn taking". We discuss this model, as it referenced in virtually all turn-taking literature. Authors either implement it (Kronlid (2006), Bohus and Horvitz (2010)) or they explicitly depart from this model (Thorisson (2002)).

SSJ starts with a thorough observation of the dynamics of a conversation. Their "grossly apparent facts [of] any conversation" are very astute (pp. 700-701):

1 Speaker-change recurs, or at least occurs.
2 Overwhelmingly, one party talks at a time.
3 Occurrences of more than one speaker at a time are common, but brief.
4 Transitions (from one turn to a next) with no gap and no overlap are common. Together with transitions characterized by slight gap or slight overlap, they make up the vast majority of transitions.
5 Turn order is not fixed, but varies.
6 Turn size is not fixed, but varies.
7 Length of conversation is not specified in advance.
8 What parties say is not specified in advance.
9 Relative distribution of turns is not specified in advance.
10 Number of parties can vary.

11 Talk can be continuous or discontinuous.

12 Turn allocation techniques are obviously used. A current speaker may select a next speaker (as when he addresses a question to another party); or parties may self-select in starting to talk.

13 Various ‘turn constructional units’ are employed; e.g., turns can be projectedly ’one word long’, or they can be sentential in length.

14 Repair mechanisms exist for dealing with turn-taking errors and violations; e.g., if two parties find themselves talking at the same time, one of them will stop prematurely, thus repairing the trouble.

In the SSJ turn-taking model, rules are defined with the main goal to achieve smooth and (close to) “one-at-a-time” speaker exchanges. In addition to the rules, there are two components necessary for the rules to work on: turn-constructional and turn-allocation components. The turn-constructional component describes what a turn can consist of. This entails ‘various unit-types with which a speaker may set out to construct a turn.’ These ‘allow a projection of the unit-type under way’. “The speaker is initially entitled, in having a turn, to one such unit. The first possible completion of a first such unit constitutes an initial transition relevance place.”(p.703). The turn-allocation component concerns the selection of the next speaker, either by self selection of by current speaker allocation. The rules are defined as follows (p.704):

1 For any turn, at the initial transition-relevance place of an initial turn-constructional unit:

   (a) If the turn-so-far is so constructed as to involve the use of a ’current speaker selects next’ technique, then the party so selected has the right and is obliged to take next turn to speak; no others have such rights or obligations, and transfer occurs at that place.

   (b) If the turn-so-far is so constructed as not to involve the use or a ’current speaker selects next’ technique, then self-selection for next speakership may, but need not, be instituted; first starter acquires rights to a turn, and transfer occurs at that place.

   (c) If the turn-so-far is so constructed as not to involve the use of a ’current speaker selects next’ technique, then current speaker may, but need not continue, unless another self-selects.

2 If, at the initial transition relevance place of an initial turn-constructional unit, neither 1a nor 1b has operated, and, following the provision of 1c, current speaker has continued, then the rule set a-c re-applies at the next transition-relevance place, until transfer is effected.

Turn taking is a "locally managed" process: it only depends on the current conversational situation, who has what conversational role, who is talking, and who is being addressed by the speaker. There is no global process that manages the conversation. Managing a conversation locally can be done in a number of ways, including, by looking at that person, or by asking that person a question Lerner (2003).

Besides common conversations there are other "speech-exchange systems" possible, such as, interviews, debates, ceremonies and meetings, trials, etc. These types can differ from conversation on a range of turn-taking parameters. For example, in meetings with a chair-person, turns are partially pre-allocated, and unallocated turns can be assigned via the use of the pre-allocated turns. In other words, the chair-person has the right to talk first, to talk after each other speaker, and they can use each turn to allocate next speakership Sacks et al. (1974),p.729. Analysis of a corpus of design meetings supports this model of a multi-layered floor structure. The floor position of the contribution is an important parameter to take into account for assessment of the social emotional value, for example politeness or dominance, of this contribution (Akker et al. (2010)).

2.3. Critiques of the turn taking model

Comments on the SSJ model of turn taking comes from various angles. In this subsection we discuss the main comments.

2.3.1. Comments on the structuralist approach

O'Connell et al. (1990) presents "a radical critique of the assumptions, concepts, methods, statistics and interpretation of data, and theories that have characterized the recent research tradition concerned with turn-taking". Their main target is the "simplest systematics" of Sacks, Schegloff, and Jefferson (1974). O'Connell et al. (1990) state:

> Instead of investigating in their own right the variables relevant to turn-taking, such as politeness and cultural norms, probabilistic speaker and hearer cues, expectations, motivations, purposes, and situational exigencies, the turn-taking research tradition has

> introduced a confusing array of purely formalistic terms such as signals, rules, devices, procedures, and systems under the general aegis of the 'turn-taking procedure'.

Where Sacks et al. (1974) claim that "the existence of organized turn-taking is something that the data of conversation have made increasingly plain" (p.699), Cowley (1998) argues that "for methodological reasons, Sacks et al. presupposed that conversations are sequences of turns" and that there is no evidence that such a mechanism as a turn-taking device exists.

We believe that the difference between Sacks et al. (1974) and critics as Cowley (1998) and O'Connell et al. (1990) is that where the former try to come up with a model that tries to describe the conversational outcomes as they are produced by the interactants, (the sentences, phrases, the turns in the sequential order in which they have been produced), the latter are more interested in the utterances, the gestures, the micro physics of the subtle signals and processes that "underly" the dialogical behavior. There are difference in the way the rules are interpreted: as descriptive, describing the "normal" way social agents behave, or as normative or even as rules that agents follows as if they were computer programs. Below we will come back on the status of the "one-at-a-time-rule".

2.3.2. Clark and Allwood's comments on the notion of "turn"

There are concerns with the validity of the construct turn, and more in general with the adequacy of the definition of the basic terms Sacks et al. (1974) use in describing their system. Clark (1996) points at the fact that the notions of turn and transition relevant place ask for a more fundamental explanation since they are defined in a circular way. In his popular book "Using Language", Herbert Clark presents his account of turn taking in conversations within his general theory of joint activity. "The placement of speech and other actions in conversation really emerges from the way people try to advance joint activities." (Clark (1996), p.327.) Clark sees SSJ's rules explained by his multilevel theory of joint actions.

> The participants in a joint activity work hard to get closure at all levels of talk - execution and attention, presentation and identification, meaning and understanding, projection and uptake. What emerges is a set of procedures that determine who speaks and acts when. These in turn account for Sacks et al.'s turn-allocation rules. Further, participants contribute to discourse in

> two phases - a presentation phase and an acceptance phase. This process determines Sacks et al.'s turn-constructional units. The two phases of contributing account for who speaks when in ways that go beyond the turn-allocation rules (adapted from Clark (1996), p.328-329.)

Clark's theory provides an explanation of *turn-taking as an emergent phenomenon*. He states that "There is no evidence that people try to preserve turns per se." (p.329). This means that turn taking rules might not be as fixed and clean as the model by Sacks et al. (1974) seems to imply.

In addition to Clarks joint action model, Jens Allwood proposes the notion of *contribution*. He suggests to analyze the different statuses that a contribution has in the context of a collaborative activity (Allwood (2000)):

> The concept of "turn" as originally put forth in Sacks, Schegloff and Jefferson (1974) can be said to be a combination of the notions of "utterance", "sentence" and "speech act" with the notions of "right to speak", "holding the floor" and "having an audience". In some cases, these notions coincide, in others they don't, which, for example, leads to difficulties in deciding whether a given contribution is a turn or not. Rather than leaving the interpretation of what a turn is open in this way, it would be preferable to connect speaker contributions analytically with a bundle of features constituted by the above mentioned concepts and admit that all of them do not always coincide.

This means that, according to Allwood (2000), a turn is a derived concept. The more basic concepts in his theory are activity and contribution.

2.3.3. Is "turn" culturally biased?

Several authors have pointed at cultural differences in turn-taking behavior. Berry (1994) found that Spanish and American people have different interpretations of the other's turn-taking behaviors which leads to misunderstanding of each others stance towards the other. Kilpatrick (1986) reports that his Puerto Rican students had the opinion that there is no such rule as Schegloff's "one party at a time" for English conversation. They felt that in Puerto Rican conversations "everyone talks at once". An analysis of recorded conversations among Puerto Rican students revealed that 90% of the speech was indeed simultaneous. This makes the definition of turn as the speech of one speaker separated by the speech of another[9] not very usefull. Kilpatrick sees turn as "a recognized speech utterance" a definition that was suggested to him by Dale Russel. "Recognized" means recognized by a hearer. You may say that a contribution is only a contribution if both parties (sender and receiver) see it that way. In fact this is not that far from Schegloff's idea of turn as an "interactive achievement".

The one-at-a-time rule is often seen as a rule of good conduct and of polite behavior. The contrast in conversational cultures makes it difficult to relate simultaneous speech to politeness or impoliteness. Puerto Rican students associated simultaneous speech with politeness, not with being rude. It seems that "turn" as conceptualized in linguistic and computational American/European research circles has a cultural bias.

Where the anthropological literature reports significant cultural differences in the timing of turn-taking in ordinary conversation (substantial overlap and a preference for simultaneous speech -see e.g. Wieland (1991)- or long pauses in between turns), recent research that test these claims shows that there are striking universals. At least, in the underlying pattern of response latency in natural conversation. Stivers et al. (2009) looked at the situations where a yes-no question was asked followed by an answer. Response time was defined as the time elapsed between the end of the question turn and the beginning of the response turn. Using a worldwide sample of 10 languages drawn from traditional indigenous communities to major world languages, Stivers et al. show that all of the languages tested provide clear evidence for a general avoidance of overlapping talk and a minimization of silence between conversational turns in these situations. They do find differences across the languages in the average gap between the turns, within a range of 250 ms from the cross-language mean. They suggest that a natural sensitivity to these tempo differences leads to a subjective perception of dramatic or even fundamental differences as offered in ethnographic reports of conversational style. There is more in conversations than question answer pairs however and the question remains: are there culturally variable turn-taking systems? Moreover, in question answering we typically have a situation where the addressee has a task (to answer) that requires the fulfillment of the questioner's

[9] The proposed ISO definition of turn is *a stretch of communicative activity produced by one participant who occupies the speaker role, bounded by periods where another participant occupies the speaker role.* (ISO/DIS 24617-2(E))

task, to state the question.

In American and Western sociolinguistic studies in the 70s and 80s gender differences in conversations were a hot topic. It appeared that females were more interrupted by males than vice versa. Sometimes this is seen as a sign of dominance of the male partner over the female (Beattie et al. (1982)). Some authors report that female conversations show more simultaneous speech than male conversations (see for example Edelsky (1981)). On the other hand, based on an analysis of a subset of Dutch telephone conversation, Louis ten Bosch (2005) reports that male-male dialogues show a higher proportion of overlapping turns than female-female dialogues. Other factors that may influence conversational turn-taking behavior are personality and status differences between interlocutors (Beattie (1981)).

There is hardly any agreement on what cultural parameters should be included in models for artificial conversational agents in order that they can show culturally colored behavior. Also there seems to be no agreement concerning gender differences in conversational behavior.

2.3.4. Thorisson et al.'s position

Thorisson (2002) argues that what misses in the SSJ model is how conversational partners recognize the turn constructional units. We have seen that Duncan (1972b) proposed the existence of "cues" for turn signalling. "Such cues are generated by interlocutors for the purpose of "signaling" to each other the state of the dialogue, such as whether they want the other to take the turn, whether they want to keep the turn, etc." Thorisson et al. claims that Duncans cues are simply the features missing from the SSJ model: the features that are used to identify the turn-constructional units, and their boundaries. Another comment from Thorisson et al. is that the SSJ model does not take into account "the internal state of cognitive processing of the participants". This "internal state" (this is the "state of mind" that Yngve proposed to model, see section 2) clearly also affects the way participants in a conversation respond to the cues in the dialogue. This "state of mind" of the conversational agent will be a core module in the architecture for natural interacting conversational agents.

2.3.5. Emotions and turn-taking

Emotions are one of the kinds of "underlying mechanisms" that play in conversations and that affect, for example, if a listener takes turn. Heylen et al. (2011b) argued that taking a more individualistic view on turn taking

-in contrast to a more conversational view from which the SSJ model is formulated- might yield interesting insights. “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.” (p.329). What are these reasons for (not) speaking? And, how do emotions influences the cognitive deliberation between the various reasons? Moreover, also intentions and motives are triggered and modulated by emotions. For example, when someone offended you, you might decide not to speak with the offender anymore. Or if you are engaged in an enthusiastic conversation, you might recurrently interrupt the other speaker in your enthusiasm.

2.3.6. Conclusion

The impressive body of work by Sacks et al. (1974) and other conversation analysts that worked on the turn-taking theory, gives a detailed description of things that can occur in a conversation. These observations are, in our opinion, falsely made into rules that govern conversation. To try and describe the dynamics of a conversation in ’simple systematics’ is brave and impressive, but has a flaw. Such simple systematics cannot describe the wide array of possible conversation types (e.g. ranging from small talk among siblings to a formal ritual like a marriage ceremony). To describe this wide array of different conversations in simple rules means that, either the rules are very general (e.g. in a conversation speakers alternate), or they are very detailed and only apply to some types of conversation (e.g. there is no overlap between speakers). For this last example, it seems clear that it does not hold for small talk among siblings, it does however, hold true for a strict ritual such as a marriage ceremony. Therefore, we suggest that observations of conversations by conversational analysts (as e.g. Sacks et al. (1974)) remain what they are, observations of conversations as they are produced. They are invaluable in naming the things that occur in conversation. In our opinion, we should work towards a model that lets the conversational dynamics emerge from an underlying system. This system *should* be able to produce the observed conversational behavior, but it should do so without having to represent the simplest systematic rules explicitly as a program for the agent to follow. This seems the only viable way to have a system that is sufficiently dynamic to represent all the various forms of conversation that exist. Also, it is the only way to prevent ending up in a race to include all the possible conversational exceptions that are possible, including those that are not yet described. The question now is, what might an underlying system for conversation (and turn taking) look like?

Turn taking seems to be *constitutive* for any multi-party collaborative activity, including conversation or dialogue: without it there is no conversation. The very concept of having a dialogue implies *in a logical sense* that those who participate in it *at least have the intention* to give the other space to talk to pay respect to his words. This seems to be consistent with Schegloff's view when he says:

> "To take one-at-a-time" to be a basic design feature in participants' co-construction of talk-in-interaction is not to assert that it is invariable achieved. If some design feature of ANY project, pursued through an organization of practices, fails to be achieved on some occasion (or even on many occasions), this is not prima facie evidence that it is not a design feature to which participants orient in the course of its production." Schegloff (2000).

The very fact that perceptually speakers speak sometimes "out of turn" presupposes a working notion of turn, as a space in time where the attention is at one of the partners. It refers to turn taking rules that serve a practical and fair distribution of the available space and time. In debates participants repeatedly remind each other of these social rules of good conduct, for example by saying "*may I finish my turn please*", or the like. What is also clear is that not all overlapping talk is problematic. But sometimes speakers talking at the same time apply some repair mechanism (Schegloff, 2000). One of the capabilities a conversational agent must have is to detect the different causes and emotional values of overlapping talk and to respond to this in a reasonable or emotional way that fits his character and the situation. We will come to the emotional capabilities of conversational turn taking agent when we discuss emotions and architectures for these type of agents.

3. Architectures for natural interaction

Dialogue systems either completely control when the user can perform what types of actions, or they give the user more freedom. In the former case the synchronisation of system and user actions is already established before the interaction starts. The user has to know the interaction protocol and it is explicitly signalled who has the turn and what actions can be performed by the user. In the latter case the system is essentially a-synchronous: while the system acts, the user acts. Thus synchronisation between user and system has to be established dynamically (it needs to be "locally managed"). This has to be learned.

Most dialogue systems are turn based, i.e. they have some notion of turn unit and they assume a fixed set of signals or cues that signal turn-taking and turn yielding. In recent spoken dialogue systems there are separate modules for topic management and for turn-taking management. The turn-taking modules of these systems are often based on the SSJ turn-taking model. A second type of systems is build on the idea of "continuous interaction". We will discuss pros and cons of both types of systems. We propose to consider the agent's attention dynamics as an underlying mechanism that shows in the agent's emergent turn-taking behavior.

3.1. Turn based architectures

The SSJ model Sacks et al. (1974) is the most popular model for turn-taking referred to by people working in the field of computational dialogue systems and social agents. Examples are Dan Bohus and Eric Horvitz' work on multi-party turn-taking Bohus and Horvitz (2010) and Kronlid (2008). See also Thórisson et al. (2010) and Traum's Mission Rehearsal Exercise (MRE) system reported in Traum et al. (2008).

Harel's statecharts (Harel (1987)), the basis for a W3C proposal for SCXML (State Chart eXtended Markup Language), are used by several authors to specify turn-taking models. Sometimes they are used for other modules of dialogue systems as well (see e.g. Heylen et al. (2011a)). Extensions of this formalism were introduced in Kronlid and Lager (2007). Extended SCXML comes close to the Information State Update specification languages such as DIPPER (Bos et al. (2003)) and Flipper (ter Maat and Heylen (2011)). Raux and Eskenazi (2009) presents the turn-taking model of the Carnegie Mellon University Dialogue System using Harel's state charts.

Kronlid's turn manager implements the SSJ system seen as a system (of "guide lines") that tells the agents who has the right to speak. Every agent has his own turn manager and dialog manager. There is an EventModule that signals a set of relevant events:

- speaker X starts speaking
- speaker X stops speaking

- speaker X will (probably) stop speaking in D units
- speaker X is addressed, i.e. X has been selected as next speaker by some other dialog partner

Similar types of events are used by the turn manager of Bohus and Horvitz (2010). The turn manager emits one of the following events:

- freeTRP: anyone may self-select
- myTRP: I am selected as next speaker
- otherTRP: someone else is selected as next speaker
- noTRP: TRP canceled
- overlap: speech is overlapping
- overlapResolved: speech is no more overlapping

The turn manager has three (parallel) charts. Figure 1 shows the three part state chart from Kronlid (2006).

The Outside chart deals with the other's states: are they speaking or silent? The Inside chart deals with the relation between the agent self and the other agents. The TRP chart deals with signaling TRPs. The TRP predictor needed to predict TRPs is assumed to exist but not explained. It says how many units it lasts until a TRP will arrive. This is the hard part of the turn manager: how to predict coming TRPs in a reliable way? Prosodic cues or a list of possible sentences are used for predicting end of turns.

There is discussion in the literature about the role of the state chart for the agent (see Raux and Eskenazi (2009)). For Kronlid the state chart is used to restrict the behavior of the agent. When some event occurs and the agent model is in a state where the event is not permitted then the agent remains in that state and the event is not “allowed”. This is the original idea of the state chart, to specify the possible traces of the system. There is however another interpretation or use of the state chart for agent modeling. In this alternative view the state are used as states of mind of the agent. They are not primarily used to restrict the behavior of the agent, for example to say what can happen or not happen in a given conversational situation, but to encode the various states so that the agent can act in an appropriate way.

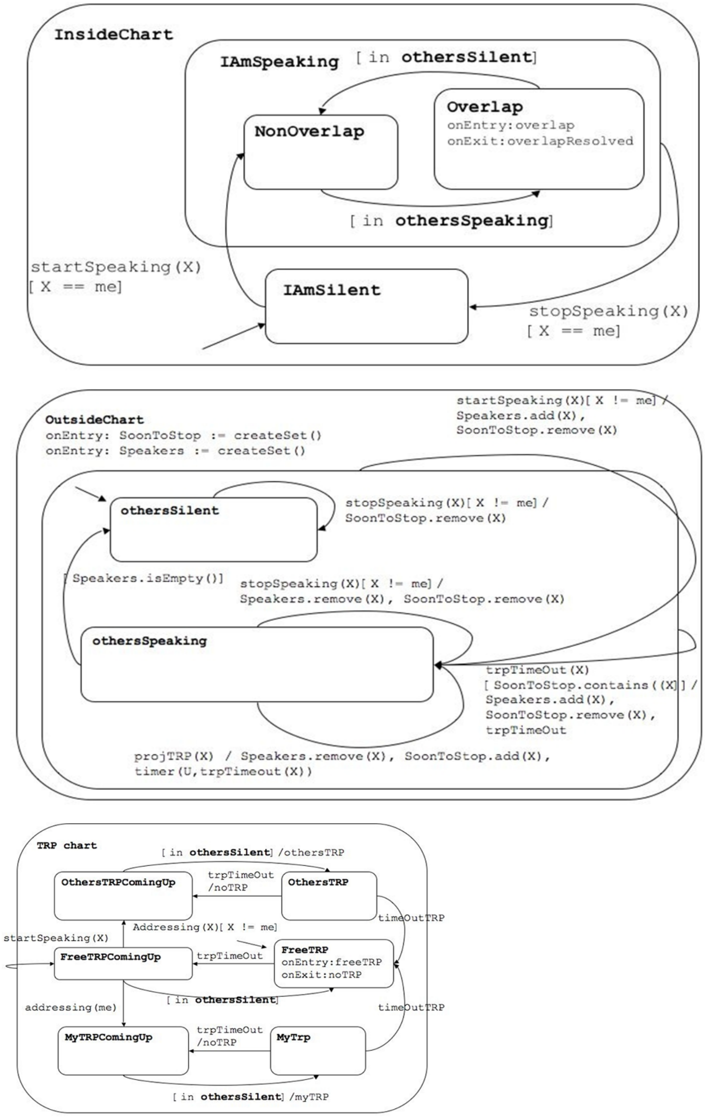


Figure 1: Three parts of the turn-taking model: Inside, Outside, and TRP chart. (from: Kronlid (2006))

This is how we will use the state chart when we use it to model the effect of emotions on the internal state of the agent. Harel state charts also allows synchronisation between processes.

Delays caused by internal processing time may cause errors. Consider the fragment in Fig. 2 taken from a dialogue between a user and the flight reservation developed at CMU Raux and Eskenazi (2007). Imagine that utterance (4) was spoken by the user not after (3) but after (2). An asynchronous Dialog Manager (DM) would still, erroneously, interpret it in state (3), as an answer to the yes/no question. However, given its timing, utterance (4) would better be interpreted as a backchannel response to the implicit confirmation (2). The second issue with asynchronous DMs is that because the DM is on hold while waiting for user responses, no execution can occur until either the user responds or a timeout is triggered. During those waiting phases, the DM cannot handle non-conversational events, which could have conversational consequences (e.g. the system might need to inform the user of a change in the real world). To address these issues, we introduce the concept of conversational floor into the execution module of the DM. The floor is an additional dialogue state variable that can take three values: user, system, and free. The value of the floor is not decided by the DM but acquired from lower level modules. Each action that the DM can plan has two markers: one indicates the value(s) in which the floor can be for this action to be executed; the other indicates the value of the floor after the execution of the action is completed. Typically, conversational acts require the floor to be free, with the exception of backchannel conversational acts and interruptions. Non-conversational actions (e.g. interacting with a backend database) also do not have floor requirements. In terms of floor transitions, the general behavior is for the floor to become User after questions and Free after statements. The DM only executes actions whose floor requirements are satisfied. When the floor is either User or System, the DM is still able to accept events, update the dialogue state, perform planning, and execute non-floor requiring actions. Both floor transitions and dialogue state updates are triggered by events from the Intermediate Layer, i.e. they reflect changes in the real world precisely when they occur. This allows the DM to interpret events, including interruptions and backchannels, in the right context. Through floor and state update events, the execution module of the DM is thus synchronized with the real-world dialogue. The combination of an asynchronous planning module with a synchronous execution module is the essence of a semi-synchronous dialogue manager.

User: I want to go to Boston. (1) System: Going to Boston. (2) System: Do you need a return trip? (3) User: Yes. (4)

Figure 2: Extract from a dialogue in the flight reservation domain.(from:Raux and Eskenazi (2007))

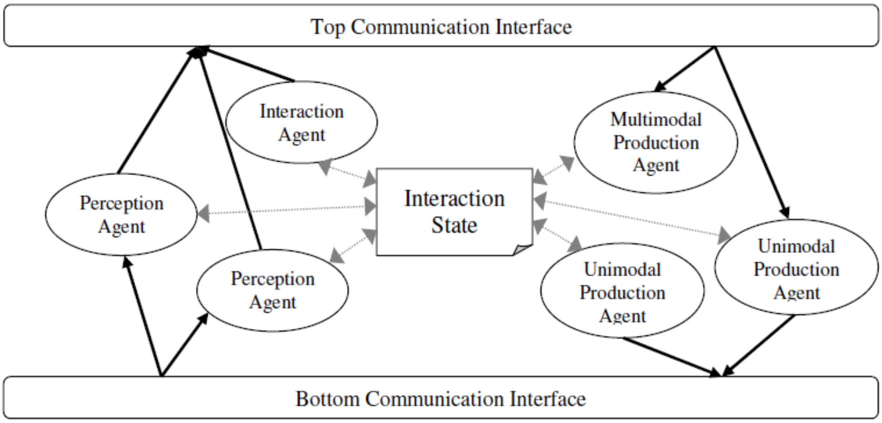

Top Communication Interface
Interaction Agent
Multimodal Production Agent
Perception Agent
Interaction State
Perception Agent
Unimodal Production Agent
Unimodal Production Agent
Bottom Communication Interface

Figure 3: Two layer model with floor state and interaction manager (from Raux and Eskenazi (2007))

Most spoken dialogue systems were build for interacting with one single human interactant. The selection of the next speaker is an issue in multi-party interaction. (See Bohus and Horvitz (2010), Bohus and Horvitz (2011), Traum et al. (2008), Kronlid (2008), Traum and Rickel (2002)).

Thórisson et al. (2010) describes an extension of the Ymir Turn Taking Model (YTTM) for multiparty turn-taking, adding to the YTTM existing functionalities the ability to model multiple speakers engaged in a "polite" cooperative dialogue. The authors support the view that turn-taking is an indirect result of the many mechanisms at play in dialogue, in particular their complex interaction and effects of limitations of realtime cognitive capabilities. "The best way to capture the operation of these many interacting mental functions in dialogue is to try to model dialogue as a fairly complete cognitive system, at a relatively fine level of detail." In the multi-party YTTM each conversation participant has an individual context model, updated with decisions from its internal deciders and input from its perception modules. Perceptions include a list of all conversation participants, who is talking, who is "looking at me" (for any given agent) and who is requesting turn at each given time. Each participant also has configuration for *urge-to-speak*; probability that another participant wants to talk (based on perceptions of their actions), the speed at which urge-to-speak rises (modeling a type of *impatience* for getting the turn), and the yield tolerance when someone else wants it (while he has turn). For any given participant, its perception of the gaze behaviors of others determines in part whether it is possible to take turn *politely*. All agents in the experimental set-up were set to be polite and collaborative. A mechanism that relates an emotional state of the agent to the variables that influence the turn-taking is not present in the model.

3.2. Architecture for continuous interaction

Many spoken dialogue systems work on a strict turn by turn basis. Building conversational agents that are able to listen while speaking, and that allow, what is sometimes called, "continuous interaction", is an active research topic (e.g. Reidsma et al. (2011), and Wang et al. (2011)). The idea of "continuous interaction" in contrast to turn based interaction, is that there is no static design time decided unit (temporal or behavioral) of interaction. On the contrary, in principle a participant can contribute something to the interaction at any time. It is up to the participants to pick up the relevant bits and pieces and to react on it. Attentive speaking and active listening require that a Virtual Human be capable of simultaneous interpretation and generation of communicative behavior. Moreover, “a Virtual Human should be able to signal its attitude and attention while it is listening to its interaction partner, and be able to attend to its interaction partner while it is speaking and modify its communicative behavior on-the-fly based on what it observes in the behavior of its partner” (Reidsma et al. 2011, p.97).

In order to realize a fluent conversation in which turn-taking is “locally managed”, some technical challenges arise. First, input from the user should be processed as soon as it is produced. This is necessary to respond verbally and non-verbally without delays, since delays disturb the conversational flow. Clark and Krych (2004) claim that speakers make their adaptations while producing their utterance almost instantly, typically initiating them within half a second of the opportunity arising. Second, the system should carefully synchronize between the various inputs it receives and the speech it generates. In interaction, and in conversations in particular, temporal order, as well as pauses, carries meaning. The systems must know the time the addressee received his words and the time that the addressee started to speak. Third, the agent should be able to distinguish the sound that comes from its own text-to-speech system from the sound that comes from the interactant. Incorporating emotions on a reactive level requires a system to solve these real-time issues. It requires models of social agents that show emotion in turn-taking and that can cope with emotionally laden overlapping talk, non-verbal contributions, and pauses by their interlocutors.

In monitoring the attentive state of the conversational partner his gaze behavior plays a role. First people gaze at something to see that something because they are interested in it. Gaze and in general focus of attention may also be a response to others’ gaze behavior. Gaze at other people is a social act with an emotional value and can be face threatening. People often avoid mutual gaze because of this emotional value of closeness or intimacy. Many authors point at the role of gaze in face-to-face conversation (see e.g. Novick et al. (1996)). People engaged in conversation may look at one another to monitor listener acceptance and understanding, to signal attention and interest, and to coordinate turn-taking. Conversely, people look away to concentrate on complex cognitive tasks. Beattie (1981) found that the role that gaze plays in turn-taking depends on context. When the overall level of gaze is low, as in conversations between strangers or when the discussion topic imposes a high cognitive load on the conversants, gaze plays a more significant role. Novick et al. (1996) explores the role of gaze in coordinating turn-taking in mixed-initiative conversation and specifically how gaze indicators might be usefully modeled in computational dialogue systems.

3.2.1. Anticipatory attention

The Ymir Turn Taking Model of Thorisson et al. Thorisson (2002) has three layers, each with their own processes running with different priorities and frequencies, for generating feedback to perceived signals. The highest priority layer is the Reactive Layer. It is concerned with behaviors that have perceive-act cycles shorter than 1 second. Reactive actions, like "looking away when you believe its your turn to speak" or gazing at objects mentioned by the presenter, belong in this Reactive Layer. The second layer is the Process Control Layer. It includes mental activities like starts and stops, interrupts; everything that has to do with the process of the dialogue task. The perceive-act cycle of such events typically lie between a half and 2 seconds (p.183). Together these two layers contain the mechanisms of dialogue management, as well as psychosocial dialogue skills. The lowest-priority layer, the Content Layer, is where the "topic" of the conversation is processed.

The Ymir Turntaking Model (YTTM) has expanded into "a broad computational model of conversational skills". However, there is no explicit attention for the inter-dependencies between turn management and the emotional state of the agent. This is also true for the later versions (Thórisson et al. (2010)). This lack of attention for combined turn-taking and emotion is an omission.

We have seen that synchronisation between autonomous processes running in their own time is a key aspect of communication. If some process A depends on the outcome of some other process B this implies that A should wait for B. A response can be given only when the request is expressed. This is the *information theoretical rationale* behind the turn-taking model. When B only needs half a word to understand what A will ask him, B can answer before A has produced the complete question. Various factors determine if B actually does that. Social relations and cultural habits play a role here. Synchronisation requires attentiveness or being sensitive or open toward the other. This depends on limited resources. People are not attentive to everything that happens all the time. It is thus quite natural that a VH while speaking does not continuously attend the actions and expressions of its addressees. A natural VH has a limited resource and his behavior will depend on the available resources. Resource (in particular energy) management and attention management are thus necessary parts of a conversational agent. One crucial factor that directs or suppresses attention and energy is emotion.

4. Emotions for virtual humans

If artificial (human-like) companions should be able to display behaviours and skills similar to human companions they should meet the following six requirements, often cited in the literature (Heylen et al. (2011a)). Social robots and agents should be able to:

1. Express and perceive emotions
2. Communicate with high-level dialogue
3. Learn and recognize models of other agents
4. Use natural cues (gaze, gestures, etc.)
5. Exhibit distinctive personality and character
6. Learn and develop social competencies

Emotions influence al other qualities of the companion. Computational models of emotion consider emotions in relation to cognition, perception and expression. These models need to provide answers to the following questions.

- How do we recognize others emotions and one's own emotions?
- How to generate emotionally colored behaviors and expressions?
- How to represent emotional state in technical systems?
- What is the relation between emotion and perception and cognition.
- What is the role of emotion in the selection of behavior?
- How does emotion influence attention?

The impression we have from the literature is that what authors call emotion is not always the same.

4.1. What do we mean by emotion?

There are a lot of different ways to define emotions. Maybe that is because there are so many different emotions and that people who study emotions focus on one of the issues above. If we say that someone is “overwhelmed by emotion”, the word emotion refers to some state of mind of a person who was impressed by some situation, something that happened, it may be a thought that caught him. Emotion is an internal force of living beings that makes them move similar to motivation, but without the cognitive and reasonable connotation that motivations have. Frijda (1986) calls them “action tendencies”. Inhibitions may prevent that we immediately follow these dispositions. Emotions come and go in many different guises and flavors. Sometimes emotions are seen as irrational, which has a negative connotation, but we also understand the function they have in our live. Cowie et al. (2011) in the collection Petta et al. (2011) gives a recent overview of the uses of the word emotion. The website of EmotionML is an entry for a lot of lists of emotion terms.[10]

Social emotions play in human social acting, acting that is directed to others and directed by others. The display of emotions is a social act. Marinetti et al. (2011) is about emotions in social interaction.

4.2. Do ECAs need emotions?

Can ECAs have emotions? What do we mean by “emotion” if we assign emotions to technical systems? Becker et al. (2007) presents motives for the integration of emotions as integral parts of an agents cognitive architecture. They distinguishes between “primary” and “secondary” emotions as originating from different levels of their architecture. Primary emotions are elicited as an immediate response to a stimulus, whereas secondary emotions are the product of cognitive processing. Primary emotions are understood as ontogenetically earlier types of emotions and they lead to basic behavioral response tendencies, which are closely related to “flight-or-fight” behaviors. In contrast, secondary emotions like “relief” or “hope” involve higher level cognitive processing based on memory and the ability to evaluate preferences over outcomes and expectations.

[10] A Working Draft of ”Emotion Markup Language 1.0”, published on 7 April 2011 can be found at www.w3.org/TR/emotionml/

Sloman et al. (2003) argues that emotions emerge from complex system behavior. They are the result of many interacting forces. They cannot be added to a system by incorporating an "emotion module".

4.3. Emotions in conversations?

What kind of emotions do we see in conversations that show in turn taking behavior? This turns out to be a hard question. There have been perception studies after the impression that turn-taking style makes on human observers. One of the first studies that addresses turn-taking and emotional feedback, produced by an embodied conversational agent, is reported in Cassell and Thórisson (1999). This perception study concerns the importance of two types of feedback: envelop feedback and emotional feedback, for the effectiveness of an "interactive computer character" when added to content feedback, such as answering a question or responding to a request. Emotional feedback is, for example, shown by scrunched eyebrows to indicate puzzlement, or a smile and raised eyebrows to indicate happiness. Envelop feedback are verbal and non-verbal backchannels and gaze behaviors produced by the agent in response to the user's communicative actions. They are related to the conversational process. A user perception study confirms the authors hypothesis that, for effectiveness of the communication, envelop feedback in combination with content feedback is more important than emotional feedback in combination with content feedback. Effectiveness is measured by the relative number of user contributions, hesitations and overlaps. Cassell and Thórisson (1999) argue that the smoother conversation is a result of users being able to apply human-human conversational knowledge to the interaction and keep track of the process of the conversation. Cassell and Thórisson (1999) used, an early version of, the Ymir model for their agent.

Humans assign different personality traits to ECAs that follow different turn-taking regimes, as was shown by ter Maat (2011). He describes four ECAs that have the goal of trying to keep the user talking. The ECAs have different personalities (sad, happy, aggressive, and pragmatic) and try to get the user in their state. For turn taking this meant that the aggressive ECA took the turn aggressively, i.e., quickly after a pause onset and often even interrupting the user. The happy ECA took the turn in a similar manner, quickly after the user stopped speaking. The sad ECA waited a moment before it started speaking. User perception was investigated for the different turn taking strategies. Quick turn taking and interrupting ECAs were perceived as more negative, and were thought to have a strong assertive personality. ECAs taking the turn after a pause were perceived as more agreeable, less assertive and users developed more rapport with them. We feel it is important to remark that there is a wide array of research on the perception of turn taking, see for example: Goldberg (1990) and Robinson and Reis (1989).

It should be noted, that in the experiments by ter Maat (2011) and ter Maat and Heylen (2009), the agents turn-taking behavior is not controlled by a dynamic model of the agents affective state itself. The model behind their agent is fairly superficial. In order to build virtual humans that behave in a believable way, we also need to model the emotional and social mechanisms behind the behaviors from which turn-taking emerges. Turn taking behavior should be the result of the autonomous actions of interacting agents.

5. Computational models of emotion

Humans have emotional capabilities. They can experience emotions, they can express emotions, they can act in a certain way in response to their emotional experiences in certain circumstances. Humans can also assign an emotional state to other humans. Humans do involuntarily express emotions they experience, but they sometimes display emotions voluntarily to communicate their emotional experience to others. Emotions are intentional in the sense that they have an object, which is not the same as their cause. Sometimes the object of the emotions is rather specific and can be determined and known, sometimes it is rather vague. Moods are emotional states that often have less clear and specific objects. There is a relation between the object and the emotion. For example, you can only be afraid of something that is frightening for you, the object has something frightening. Emotional capabilities are to be distinguished from character traits. Someone who is capable of experience fear need not be a timorous sort of person (see Goldie (2000)). Character traits are dispositions for having thought and feelings of a certain sort. A friendly and social open character will more easily feel happy in a social conversation than an introvert character. Emotions influence the way we perceive what is going on in a certain situation as well as the way we cope with the situation. Some theories distinguish a small number of emotions as basic emotions. The involuntary immediate bodily expression in facial expressions of these emotions is often considered pan-cultural. This holds also for the recognition of the emotions by experiencing the facial expressions displayed by someone enacting the emotion.

In this section we discuss some of the prevailing computational models of emotion. The role these models have for the researcher depends on his particular interest and his research area. Computational models of dialog as well as computational models of emotions are developed in the field of social psychology and social linguistics. They are used then as theoretical constructs to improve understanding of the subject matter of these science. According to Marsella et al.(2009) the advantage is that computational models enforces to lay out the theoretical concepts in more detail. Emotion (as well as dialog) models are also developed and used in the field of human computer interaction. Here, computational models of emotion are used for understanding the emotion of the user related to the interaction with a technical system, so the system can better adjust its behavior to the user, (Picard (2003)). Third, computational models are developed in the field of artificial intelligence, for example to build robots or embodied agents. The idea is that taking emotions into account makes systems more intelligent. Systems that have emotions are taken to be better of in some situations than systems that don't have emotions. A fourth interest in computational modeling of emotions is in the field of logic and philosophy. Here the research not only aims at understanding emotion but also asks how in different disciplines emotions are understood. Philosophy asks what emotions are and what social emotions are, whereas in computational emotion psychology emotion is a basic concept. We will not delve into a detailed discussion about the concept of emotion here, but note that affect and cognition are often seen as complementary notions, whereas affect is an essential aspect of the cognitive relation we have to other objects and people. This shows most clearly in the fact that we want to know everything about the ones we feel affectively attracted to. Involvement is a cognitive as well as an affective parameter.

It goes without saying that the content of a computational model of emotion depends on the role it plays for the researcher. For example, an emotion model for a virtual character that plays a role in a virtual story or game should be such that the agent behaves in a believable way.

Here we are interested in computational models of emotion that contribute to the fabrication of agents that show natural conversational behavior. This conversational behavior will often be part of other activities that the agent takes part in, mostly in joint operation with other agents, artificial or human. Applications are tutoring systems or (interactive) story telling systems. The emotions that play in a social interaction are part of the "what is going on" in that activity. Important parameters are among others: social distance and familiarity, power and dominance, and how serious the situation is. These parameters “color” the social interaction. Conversations have different colors or moods depending on these parameters. An interrogation of a crime suspect by a police officer has a different mood than a chat between a couple of close friends. Questions that the psychologist asks to a client in a psycho-analytical session are different in flavor than the standard questions that an interviewer asks to someone he doesn’t even know. These differences have to do with the different ways that the people are involved, how important the event is for them, what their goals are, what the risks involved are, how they experience the (conversational) joint activity. What is important for the participant is how they see “the other”, how they think “the other” sees him, and how he wants that “the other” sees him. Here “the other” can be many different concrete others, actually present as a concrete individual our subject is interacting with, or some vague image of “the other”, a social conscience. In a political debate broadcasted on television the debaters show their awareness of the audience. The politician will try to show the public that he can counter the arguments put forward by his competitors. He will fight for the floor in order to gain enough space to sell his political story. But he also shows that he is willing to listen to the other and that he respects the other even if he doesn’t agree with his political view.

This brings us to the important notion of face and the relation it has with social emotions. What kind of emotions play a role in social interactions and how are these emotions related to face? Face and facework are often seen as the main ingredients of politeness and politeness strategies. (Brown and Levinson (1987) and Goffman (1967)). Interrupting the speaker is often seen as impolite behavior. Politeness is sometimes seen as “strategic conflict-avoidance”. The idea is that the basic social role of politeness is in its ability to function as a way of controlling potential aggression between interactional parties. What is polite is socially appropriate behavior and what is socially appropriate depends on the speakers social position in relation to the hearer.

If interrupting a speaker is impolite, why do people interrupt others? Because there are other emotional forces that play a role. Forces that are so strong that the resulting force makes the agent act. The subject may, overwhelmed as he is by his spontaneous act, immediately feel sorry or even ashamed for it and apologize for his behavior.

5.1. Cognitive emotion theories

The emotion model proposed by Ortony, Clore and Collins (Ortony et al. (1990)) has often been the basis for the integration of emotions into cognitive architectures of embodied characters. It has been suggested that it might be sufficient to integrate only ten emotion categories, five positive and five negative ones, into an agents architecture when focusing on believability and consistency of an agents behavior.

Ortony distinguishes three “emotion response tendencies” by which the effect of an emotion on the agents cognition can be characterized: facial display, information-processing and coping. Coping deals with the fact that humans are able to cope with their own emotions in a problem-focused or emotion-focused way by trying to change the situational context (problem-focused) or by trying to reappraise the emotion eliciting situation to manage their own emotions internally (emotion-focused).

Based on neurophysiological findings, Damasio distinguishes “primary” and “secondary” emotions. Primary emotions are elicited as an immediate response to a stimulus, whereas secondary emotions are the product of cognitive processing. Primary emotions are understood as ontogenetically earlier types of emotions and they lead to basic behavioral response tendencies, which are closely related to “flight-or-fight” behaviors. In contrast, secondary emotions like “relief” or “hope” involve higher level cognitive processing based on memory and the ability to evaluate preferences over outcomes and expectations.

5.2. Dimensional emotion theories

Wundt (1922) has claimed that any emotion can be characterized as a continuous progression in a three-dimensional space of connotative meaning. Several researchers, e.g. Gehm and Scherer (1988), Mehrabian (1995), have later provided statistical evidence for this assumption. Using principle component analysis they found that three dimensions are sufficient to represent the connotative meaning of emotion categories. These dimensions are commonly labeled Pleasure/Valence (P), representing the overall valence information, Arousal (A), accounting for the degree of activeness of an emotion, and Dominance/Power (D), describing the experienced control over the emotion itself or the situational context it originated from. The three-dimensional abstract space spanned by these dimensions is referred to as PAD-space.

Lang (1980) devised a picture-oriented instrument called the Self-Assessment Manikin (SAM) to directly assess the pleasure, arousal, and dominance as-

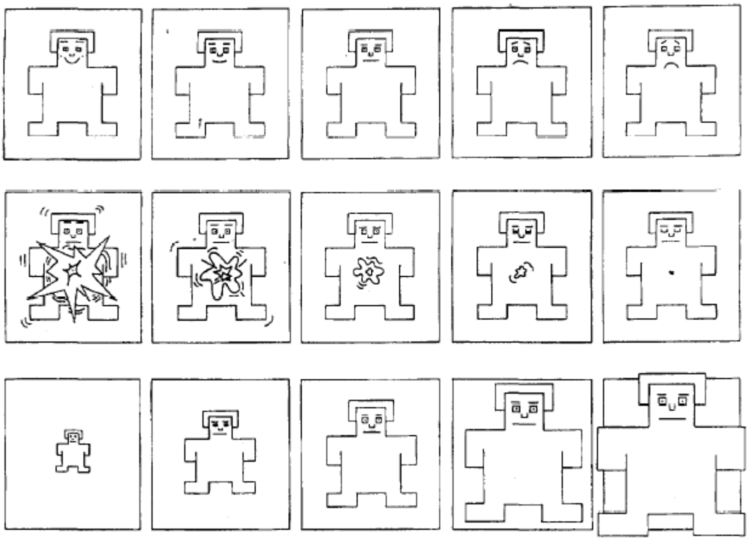


Figure 4: The Self-Assessment Manikin (SAM) used to rate the affective dimensions of valence (top panel), arousal (middle panel), and dominance (bottom panel). From Bradley and Lang (1994)

sociated in response to an object or event (Bradley and Lang (1994)). See Figure 4.

5.3. Face and politeness

The face is understood as something that is emotionally invested, and that can be not only lost, but also maintained or enhanced. Goffman (1967) defines face as "the positive social value a person effectively claims for himself by the line others assume he has taken during a particular contact. In addition, it is an image of self delineated in terms of approved social attributes(...)". The relation between politeness and face was studied in Brown and Levinson (1987). They state that every individual has two types of face, positive and negative. They define positive face as the individuals desire that her/his wants be appreciated in social interaction, and negative face as the individuals desire for freedom of action and freedom from imposition. B&L's theory assumes that most speech acts, for example requests, offers and compliments, inherently threaten either the hearers or the speakers face-wants, and that politeness is involved in redressing those face threatening acts. On the basis of these assumptions, three main strategies for performing speech acts are distinguished: positive politeness, negative politeness and off-record politeness. Positive politeness aims at supporting or enhancing the addressee's positive face, whereas negative politeness aims at softening the encroachment on the addressee's freedom of action or freedom from imposition. The third strategy, off-record politeness, means flouting one of the Gricean maxims (Grice (1975)) on the assumption that the addressee is able to infer the intended meaning. Speakers calculate the weight of their speech acts from three social variables: the perceived social distance between the hearer and the speaker, the perceived power difference between them, and the cultural ranking of the speech act. The latter is defined as the degree to which the face threatening act is perceived to be threatening within a specific culture. On the basis of the outcome of the calculation, speakers choose the appropriate type of strategy and substrategy to be employed. Next, they select the appropriate linguistic means by which to accomplish the chosen substrategy. Different linguistic structures realize specific strategic choices. For example, one of the substrategies addressed to the hearers negative face is Be conventionally indirect, and when the speaker selects this strategy for asking to pass the salt, s/he may choose the structure Could you pass the salt?" (See also Vilkki (2006) where B&L's equation of politeness theory and face theory is questioned.)

5.4. The EMA model of emotion

Marsella and Gratch (2009) present a computational model of emotion. They argue that the often used multi-level theories for appraisal complicate appraisal processes by conflating appraisal and inference. They argue that appraisal and inference are distinct processes that work on the same mental representation a person has over their relationship with and in the environment. They view process validity a key criterion for evaluating a model of emotion. Process validity, for them, means that the model captures the unfolding dynamics of emotions. To test this, the model's behavior and emotion are compared to human data for a situation where emotions occurred. A surprise/startling event (i.e. a bird flew against a window) yielded the following sequence of reactions across participants: raised eyebrows, lowered eyebrows and jaw drop, expressions suggesting relief, amusement, and compassion. Such behaviors suggest a transition of emotions from surprise, to concern for own safety, to concern for others, in response to an unexpected event. This example shows that emotions are dynamic rather than fixed. Perceptual and inferential processes alter the interpretation of the situation. These processes require time to draw inferences, and over time more information can become available or the situation can change. This all contributes to the dynamics of emotions. In other words, the emotion evolves as cognitive processes update the person-environment relationship. Authors summarize the requirements for a computational model of emotion as a process that can interpret the person-environment (or in this case, the agent-environment) relation. This interpretation of the system is characterized as a set of criteria, and specific emotions are associated with certain configurations of these criteria. Additionally, appraisal theories that underlie the EMA system posit specific appraisal dimensions that impose additional requirements (i.e., relevance, valence and intensity, future implications, blame and responsibility, power and coping potential, and coping strategies). Finally, there are specific process assumptions that Marsella and Gratch found necessary for concretizing appraisal theory into a computational model. The notion that appraisal causes emotion is rooted in the notion by Frijda (1988) as the law of situated meaning. To address the dynamics of emotional processes, EMA assumes a cycle of appraisal and re-appraisal. This means that an initial appraisal of a situation evokes cognitive and/or behavioral responses, which change the person-environment relation resulting in a cyclical relation between appraisal, coping and re-appraisal. Finally, Marsella and Gratch argue that there should be a clear distinction between inference and appraisal. They state that appraisal is a quick and shallow process that forms emotions. This occurs in parallel to multiple processes, both perceptual and cognitive, that make inferences about the person-environment relation. See figure 5.

In EMA, a blackboard-style model keeps track of the representation of the ongoing (causal) interpretation of the agent-environment relations. Knowledge in this representation is explicitly organized in past, current, and future events, whereby events are organized in a causal manner. Every event representation has a probability and a belief component (whether or not the agent beliefs this event/action is true). For actions, there is also a property of who did it, and whether or not it was intentional. This is important for things like assigning blame. Finally, there is a property that assigns the preference agents have for states. Appraisal, in EMA, is modeled as a set of continuously active feature detectors that map features of the causal interpretation into appraisal variables. All the objects in the causal interpretation of the agent-environment are appraised separately, simultaneously and automatically. The model uses several appraisal values associated with each object: relevance, perspective, desirability, likelihood, expectedness, causal attribution, controllability, and changeability.

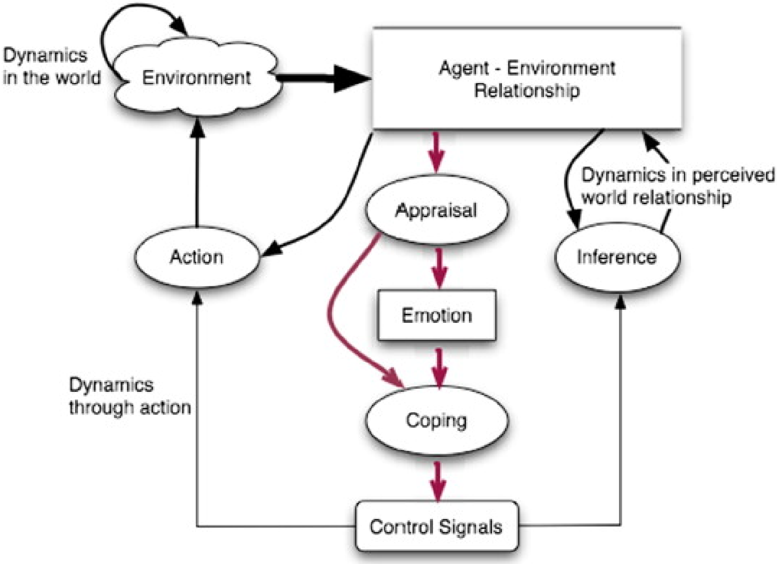
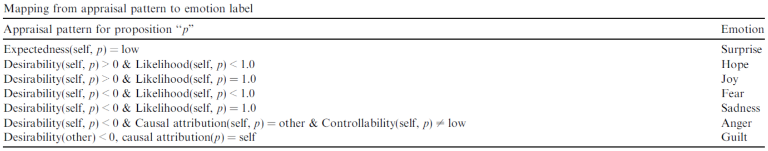


Figure 5: From Marsella and Gratch (2009)



Mapping from appraisal pattern to emotion label

Appraisal pattern for proposition "p"	Emotion
Expectedness(self, p) = low	Surprise
Desirability(self, p) > 0 & Likelihood(self, p) < 1.0	Hope
Desirability(self, p) > 0 & Likelihood(self, p) = 1.0	Joy
Desirability(self, p) < 0 & Likelihood(self, p) < 1.0	Fear
Desirability(self, p) < 0 & Likelihood(self, p) = 1.0	Sadness
Desirability(self, p) < 0 & Causal attribution(self, p) = other & Controllability(self, p) $\neq$ low	Anger
Desirability(other) < 0, causal attribution(p) = self	Guilt

Figure 6: From Marsella and Gratch (2009)

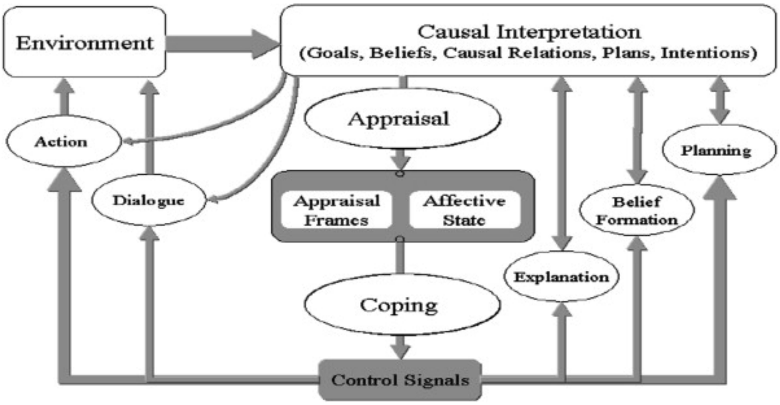


Figure 7: From Marsella and Gratch (2009)

Emotional responses are generated by linking specific appraisal states to certain emotions, see for example table in Figure 7. In EMA, these emotional labels are used as a convenient shortcut to generate emotional responses on an agent (e.g. facial expressions). In the EMA model, see Figure 6, appraisal is separate from, but closely coupled to, the perceptual, cognitive and behavioral processes. Emotion, the appraisal frame and affective states, has a central role in the model. The authors state, this is in keeping with the view of emotion as an interrupt mechanism. According to this view presented in the sixties by Herbert Simon emotion is an interruption mechanism that allows a system to respond in real time to urgent needs (see Simon (1967)).

6. Architectures for an emotional agent

We have seen that there are many different types of emotions, different computational models of emotions. There are different motivations for building emotional systems. As a consequence there are different architectures for modelling emotional agents. Here we discuss two typical models. In most systems emotion modeling is inspired by the OCC model developed by Ortony, Clore, and Collins (Ortony et al. (1990)). Approaches differ in the granularity of modeling, the mathematical machinery for computing emotions, and in the way of how the model has been implemented on a technical level.

6.1. An action selection and appraisal architecture for emotional agents

Burghouts et al. (2003) presents an architecture for an emotional agent that can be used as human computer interface. The underlying model relates emotion with cognition and behavior. The functional theory of emotions of Frijda (1986) forms the main theoretical background for the construction of the model and architecture. According this theory emotion is change in action readiness or action tendency. Figure 8 shows the architecture that models the emotion process, cognition and behavior. The boxes to the left under the label Individual list various characteristics of an individual (agent) that play a role in the emotion process, such as character and personality.

The emotion process is defined by four modules that each change the intensity of the emotions in a rule-based fashion: appraisal, activation, inhibition and self-control.

Appraisal The agent continually evaluates its environment. During this appraisal process emotions are elicited. This proces is based on the OCC model. Emotions are triggered or elicited depending on certain conditions and decay over time.

Activation In the activation stage, the emotional state is transformed to "potency values" for activated emotions, depending on habituation, expectancy and discrepancy.

Inhibition Suppression of emotions may occur due to inhibition. Inhibition is especially evident in the case of negative emotions. An individual will then try to reduce these emotions by emphasizing positive emotions.

Self-control Emotions differ with respect to the ease with which they can be controlled. It is often claimed that some hot or old emotions, such as fear, are difficult to control. Self-control is manifested in the reduction or removal of dissonances between expected behavior (based on the individuals character) and the standards one upholds (i.e. the way the agent judges the praiseworthiness of actions).

Various factors play a role in the selection of an action. For instance, how one perceives the success of behavior depends heavily on ones confidence as well as on the level of extraversion. The action selection components are rule based and determine a strategy to actually select an action.

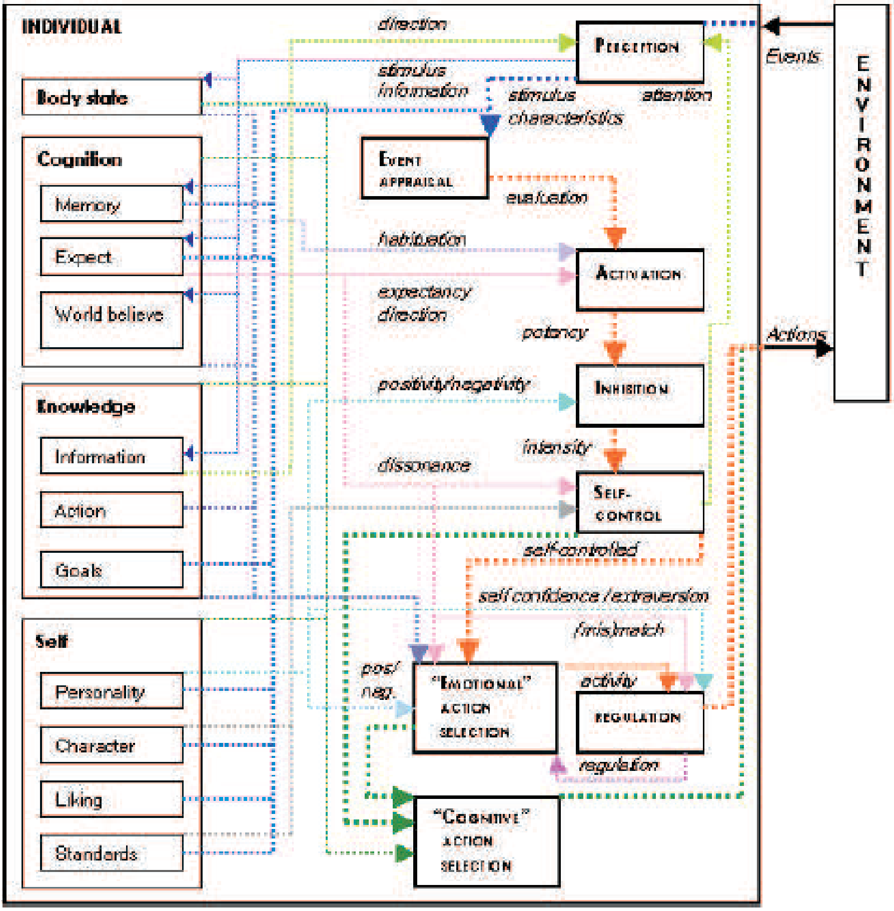


Figure 8: An action selection and appraisal architecture (from: Burghouts et al. (2003))

The architecture was tested in a prototype environment with different characters food. It has food and predators. This was set up to see how the components in the architecture could be made operational for certain agents and how this would influence the behavior of these agents living in the environment. Hero and Grumph are inhabitants with two different personalities. They differ in the way they appraise events. Hero is fairly optimistic. He is more confident and will for instance be more inclined to attack predators. He is also self-centered, extraverted and idealistic. Grumph is inclined to blame others for his misfortunes. He is easily disappointed, reproaches others more often and is introverted. The question asked in Burghouts et al. (2003) is then if "we can consider our agents behavior to be believable?" The main achievement is that the architectures generated shallow emotional behavior is believable from observation. "The apparent differences between Heros and Grumphs behavior relate to our intuition of an idealistic, extraverted, self-confident but self-centered personality, and an introvert and negative personality, respectively".

The architecture covers the complete emotion process. There is however no interaction between the emotional characters and the human user. Burghout's architecture implements a linear emotion process. In the next subsection we will consider a multi-layered model of emotion processing.

6.2. A multi-layer architecture for emotional agents

Sloman (2001) distinguishes shallow and deeper models of emotion. Within an architecture-based theory three types of emotions are distinguished: primary emotions, secondary emotions, and tertiary emotions. The taxonomy given in Ortony et al. (1990) focuses on a particular set of cognitive and motivational states and attitudes and can be accommodated within the classes of secondary and tertiary emotions. The architecture should support the system to show certain input output behavior which is perceived by humans as "emotional" of a certain type. So that we can explain the system's behavior as an act partly motivated by some specific emotion (fear, disgust, envy, shame) that we assign to the state of the machine. According to Sloman et al. (2003) "the more human-like robot emotions will emerge from the interactions of many mechanisms serving different purposes, not from a particular, dedicated *emotion mechanism*".

One of the aspects of the "primary" emotions is the immediate interrupt effect on the ongoing processing stream. This requires a level of acting that can immediately interrupt "higher" levels of deliberate behavioral processing.

For example, the speaker stops speaking immediately when he suddenly sees that someone on a short distance throws a tomato at him.[11] This is what LeDoux calls the "pre-cognitive emotion." LeDoux (1996). They take the fast (a millisecond) express route from the eye or the ear via the thalamus to the amygdala which immediately leads to action, before the neocortex has decided what is going on. (See Dalgleish (2004), p.585 for a review of neurological experiments that provide evidence for the central role of the amygdala for the emotion of fear.) According to Adolphs (2004) LeDoux follows an approach to emotion in the absence of feeling: emotion as behavior without conscience experience. Sloman has a similar approach to emotion.

In AI Sloman argues it is common to distinguish *reactive* mechanisms in which states detected by sensors (whether external or internal) immediately trigger responses (whether external or internal) from *deliberative* mechanisms in which alternative possibilities for action can be considered, categorised, evaluated, and selected or rejected. Such a deliberative mechanism is capable of "what if" reasoning. (Sloman (2001), p.11). Notice that in technical systems this distinction is in the eye of the human designer only. That is, some actions are seen as reactive response to a sensor input others are seen as the results of a process of reasoning in which multiple variables determine the process outcome. But it will be clear that also responses that are seen as reactive are described by means of "if-then" rules. From the AI perspective it is easy to forget this: the difference between the meaningful processes of the computer and the physical processes of the same thing only exist for the human who understands these processes as state changes and computations. If emotions are essentially aspects of live, as we think they are, then robots and ECAs must be seen as living beings in order to assign them an emotional state.[12] Hollnagel (2003) argues that since computers only have one system

[11] Programming languages, for example Java, support multi-threading and exception handling. The lower level proces throws an exception that is caught by the monitor of other processes, and acts as prescribed in an exception handling module.

[12] From the perspective of AI there is no difference between living systems and technical systems. They are both conceived as information processing systems. Sloman seems to have the idea that the essential (relevant) difference between these two types of systems is in the complexity of their "architectures". Emotion emerges as a result of interacting processes in a complex way. Also the difference between internal causes or motives and external causes disappear from the AI perspective. Discussions within this AI perspective that try to define various forms of "autonomy" of artificial agents, such as "social autonomy", "goal autonomy" and "executive autonomy" are bound to result in circular

of logical information processing and lack anything similar to an autonomous nervous system, “there is no way which they can be emotional or affective in the normal meaning of the words.” “The term affective computing is therefore an oxymoron.” (p.68).

Figure 9 shows an architecture for a social conversational agent from Steunebrink et al. (2008). The design shows the three different routes that emotions follow onto the effects on expression and behavior. These three routes coincide with the three layers of Sloman’s model.

As an example of a reactive and a cognitive mechanism in a conversational agent system we can think of two levels of verbal input processing. A “reactive” proces recognizes occurrences of special key words in the input that it receives from a keyword recognizer. If a certain key word is recognized this triggers a standard “emotional response”. A second, more cognitive process, recognizes complete utterances based on more in depth sentential and contextual analysis. This takes more time, but is more precise.

On a third level the system reflects on the lower levels. It perceives and evaluates the acts and effects that results from the acts caused by the emotions on the lower level. Here emotions like shame can result if the agent realizes that it has done something that it feels responsible for, but that he might not have done if he had full control over his acting. For example, the agent interrupts a speaker because he feels he should say something but as soon as he did he feels sorry and apologizes. The system can learn from experience. For example in our example of shallow versus in depth language processing the system can adapt the list of key words and associated responses based on the deeper analysis and the effect of the “emotional” response.

This requires that the turn-taking system has special states that tell the agent not only that he speaks but also what motivated or caused him to enter this state. The meta-level process that reflects on this state can generate appropriate repair behavior for example face work.

descriptions. “For having goal autonomy an agent must be endowed with *goals of its own*.” and “This is the main point: if the agent decides to do something *for its own reasons (goals)* its decision is autonomous whether it complies with others’ request or not.” (Castelfranchi and Falcone (2003), both citations p. 108, italics by RodA and MB.) These kinds of descriptions don’t clarify anything without a clear idea of what it means to do something “for its own reasons” or “to have goals of its own.”

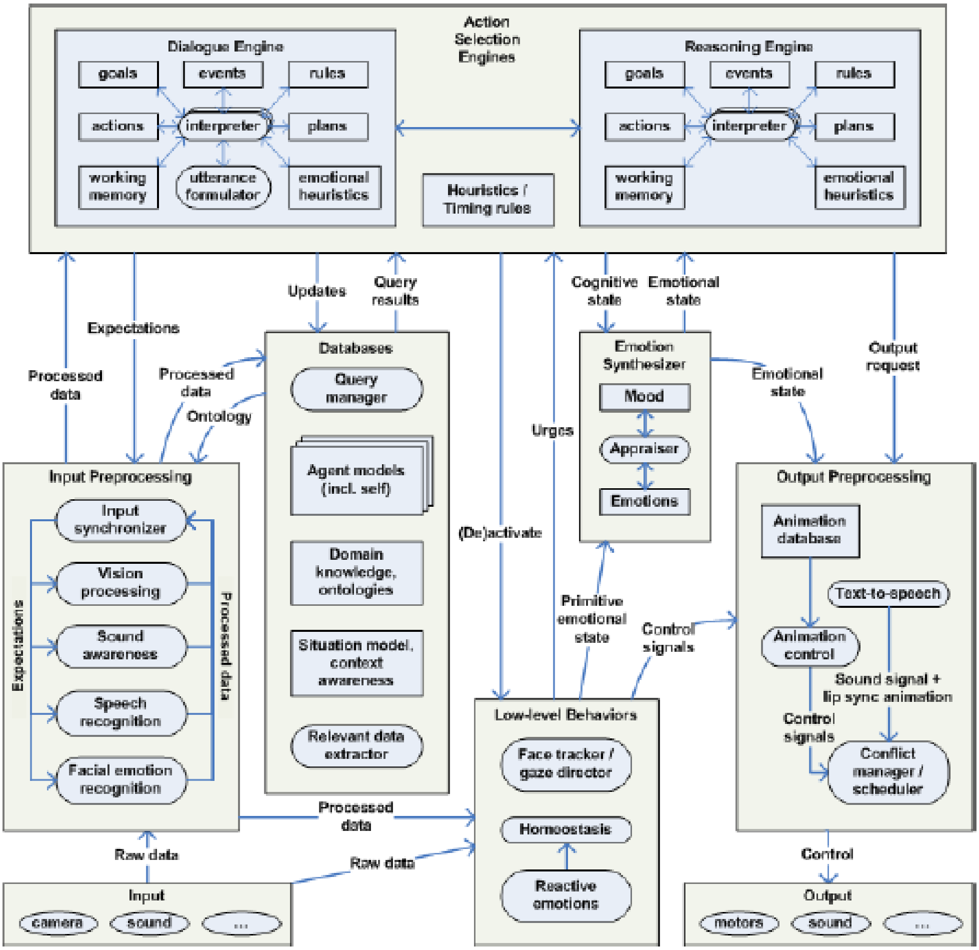


Figure 9: A generic architecture for social conversational agents with a layered emotion model (from: Steunebrink et al. (2008))

7. Sketch of an architecture for emotional turn taking

In this section, we sketch a multi-layer architecture for a conversational agent that has the capabilities that we have identified in sections 4 and 5 so that it incorporates the relations between emotions and turn-taking behavior. We recall the following capabilities.

1. The agent is able to identify emotional states of its conversational partners. In particular it can recognize those emotions and stances that are relevant in the given task and context.
2. While listening as well as while talking the agent is able to evaluate what is being said by the speaker(s) and to assess the emotional and face value of this.
3. The agent is able to predict the emotional and face value of his own acts, including what he says as well as what he does with respect to turn taking.
4. The agent shows emotional awareness on three layers. On the first layer he is continuously sensitive for low-level inputs which lead to an immediate response. On the second layer the agent acts emotionally on an event that attracts his attention. On this layer, contrary to the previous first layer there is cognitive assessment by some kind of appraisal mechanism. On the third reflective layer the agent assesses his own emotional state, mood and behaviors.

Based on our discussion about turn-taking and emotion architectures in section 3 and section 6 we propose a distributed publish subscribe blackboard architecture with typed message passing between various processors and modules. Moreover, see our conclusion in section 3 based on our analysis of the turn based model and the model of continuous interaction, the agent has an attention module containing the attentive state of the agent. The attentive state determines how the agent distributes his bounded resources as attention to the outer and inner events and processes.

The architecture that we propose is a combination of the YMIR layered architecture for dialogue agents of Thorisson (2002) and the layered emotions architecture of Sloman (2001). To realize requirements (1) and (2) the architecture has incremental streaming of "real time" input modules for audio and video. The speech recognizer and the linguistic and para-linguistic analysers provide input to cognitive layers with a timing controled by the attentive state of the agent. To realize requirement (3) the agent has a Social Emotion and Face module that assesses the face value of own and others actions. Because of (4) the proposed architecture has three layers. The main modules implement the different dimensions that need to be managed for our agent to interact with other agents which are supposed to be of a similar design.[13]

Here we elaborate shortly on the turn taking management and the way turn taking interacts with the emotional layers and the social (face) module. The state of the turn taking module encodes the notions of turn and floor as discussed in section 2. The state model has two functions. The first is to control the possible actions of the agent by means of guarded transitions as in state machines (Blanch (2006) and Harel (1987)). The second is for reflection: the agent is aware of the state it is in. Similar as in the turn taking system of Kronlid (2006) the agent has separate (super)states for his own and the other agents states. Topic management and turn taking management are connected as in the architecture of the SERA demonstrator of Heylen et al. (2011a).

The agent can take turn (in the sense of start talking) either voluntarily, based on a decision process by the dialogue managers intention planner or emotionally, based on a response on the second layer. Note that process outcomes on the first (lowest, automatic) layer will not affect conversational turn taking content wise. Events processed on this layer may affect the speech output: the agents stops speaking caused by a direct cause on the physical level. The turn state models in it state the cause of the turn state. By the reflective layer the agent can become aware that he took the turn and interrupted the speaker because of some affective response. The face module implements the required functions for the face work: the agent may for example apologize for speaking out of turn.

Given the fact that resources are limited, perceptual processing depends on the *focus of attention* of a dialogue system. The dialogue agent anticipates expected events by directing his attention on specific input channels or events. The attention focus depends on the activity the agent is involved in. In the listening role he will be focusing more on the content of the speaker, in the speaker role he will be more focusing on the typical types of listener responses. Global context features such as the general mood of the dia-

[13] That means that some level of "ritualized" interactive behavior in the community of interacting agent is assumed to be established when they interact. Note that such a "common ground" needs to be programmed, so it can be assumed "by design" when we deal with fabricated agents.

logue, the task, roles and social relations between participants influence the expectation the agent is anticipating. This implies that the attention modules receives input from the turn taking module. Apart from this controlled mechanism of attention focus anticipating expected events, there is a second "mechanism" that has a more random character. Things just happen. Sometimes they are caught by an attentive listener, sometimes they are not. That's life.

8. Conclusion

When we started this review the selection criterium for literature to read and review was that its focus was on computational models and/or architectures and/or systems about (social) emotions in relation to turn-taking in natural conversations. It turned out that there is hardly any literature that satisfies these criteria. There is a lot about computational models of emotions, there is a lot about turn-taking models but it appears that VH nowadays cannot decide about the (social) emotional value of the speech of the human speaker. Nor can they reason about their own involvement with the interaction and "decide" if they should interrupt the speaker and how to express their stance towards what is being said or to the speaker. These are capabilities that humans have and that we observe in natural conversations. We decided to review papers in both areas in computational modelling and systems.

This void in the overlap of the two fields is understandable. Emotion and conversation belong to two different research areas. But the main reason is that VHs that can show context-sensitive emotional turn-taking in conversations need real-time speech and non-verbal input processing. Good quality real-time speech recognition is the bottleneck even for task-based spoken dialogue systems. Moreover, these VHs need complex architectures that allow multi-processing and synchronisation. This holds for the "internal" processes for the different layers of activity of the agent itself, but in the first place for the essentially "autonomous" processes of the participating agents each of which run in their own time.

For building believable virtual humans that converse with human agents we need architectures that allow free turn-taking. A number of existing architectures allow the freedom to take turns whenever the situation asks for it in the eyes of the agent. Emergent natural emotional conversations implies free spontaneous turn taking. Computational models of emotion are mostly focusing on the basic emotions, hardly on social emotions. In particular face concerns and politeness issues, which play important roles in conversations, have not been taken into account in virtual human architectures in relation to turn-taking. They restrict to how politeness affects the choice of dialogue act and linguistic style. Simultaneous speech can emerge in conversations either caused by an emotion process or motivated by a deliberate cognitive process. Architectures for natural free turn-taking should be extended with modules that are able to decide what caused the simultaneous speech, that adapt the emotional state of the agent as well as a module that interacts with the turn-manager that decides when to talk and that generates an appropriate response behavior based on the emotional state. Moreover, such an agent architecture should have different modules each with their own processes for emotion handling and reflection. We have presented some initial ideas how the different layers could be modelled using hierarchical state charts.

Well designed architectures are a prerequisite for building model based embodied conversational agent behavior. But architectures don't work. Where humans can simultaneously talk and pay attention to their interlocutor's actions, filter the relevant bits and almost immediately react on that, technical real-time recognizers, such as real-time speech recognition systems, are still a bottleneck for building conversational agents that show real-time interactive behavior in which the agent reacts on what the speaker is saying. Building scripted conversations between virtual humans, similar as in the CrossTalk system (Rist et al. (2002))[14], or the web-based SCREAM system described in Prendinger and Ishizuka (2002) seems to be a good first step in developing and evaluating such extended models. They allow us to study how human observers perceive the mood and emotions that play in an interaction that is played by virtual humans.

In 1970, Yngve was one of the first who pointed at the main ingredients of a computational theory of the "state of mind" of the conversational agent. In his proposal that extends linguistic research to research in conversational behavior, Yngve makes clear that "turn", "floor" and emotional stances such as shown by "enthusiastic and animated agreement" in simultaneous back channel contributions are essential concepts in such a theory. "What remains to be done (...) is to produce a structural description of the state of mind and show explicitly how it is related to communicative behavior in its totality,

[14] www.dfki.de/crosstalk/

including the linguistic details." (Yngve (1970), p.571)). A lot has been done to accomplish this. A lot remains to be done.

Acknowledgment

This publication was supported by the Dutch national program COMMIT.

References

Adolphs, R., 2004. Could a robot have emotions? theoretical perspectives from social cognitive neuroscience, in: Arbib, M., Fellous, J. (Eds.), Who needs emotions: the brain meets the robot. Oxford University Press, pp. 9–28.

Akker, R.o.d., Heylen, D.K., 2007. Feedback loops in communication and human computing, in: Pantic, M., Pentland, A., Nijholt, A., Huang, T. (Eds.), Artifical Intelligence for Human Computing. Springer Verlag, Berlin. volume 4451 of *Lecture Notes in Computer Science*, pp. 215–233.

Akker, R.o.d., Theune, M., Truong, K.P., de Iwan, K., 2010. The organisation of floor in meetings and the relation with speaker addressee patterns, in: Proceedings of the 2nd international workshop on Social signal processing, ACM, New York, NY, USA. pp. 35–40.

Allwood, J., 2000. An activity based approach to pragmatics, in: Abduction, Belief and Context in Dialogue. Studies in Computational Pragmatics. John Benjamins Publishing Company.

Bates, J., 1994. The role of emotion in believable agents. Commun. ACM 37, 122–125.

Beattie, G., Cutler, A., Pearson, M., 1982. Why is mrs. Thatcher interrupted so often? Nature 300, 744747.

Beattie, G.W., 1981. Interruption in conversational interaction, and its relation to the sex and status of the interactants*. Linguistics 19, 15–36.

Becker, C., Kopp, S., Wachsmuth, I., 2007. Why Emotions should be Integrated into Conversational Agents. John Wiley & Sons, Ltd. pp. 49–67.

Berry, A., 1994. Spanish and american turn-taking styles: A comparative study. Pragmatics and Language Learning. Monograph Series 5, 180–90.

Bickmore, T., Picard, R., 2005. Establishing and maintaining long-term humancomputer relationships. ACM Transactions on Computer-Human Interaction 12, 293–327.

Blanch, R., 2006. Facilitating post-WIMP Interaction Programming using the Hierarchical State Machine Toolkit. Technical Report. Commune de Recherche en Informatique du plateau de saclay CNRS, Ecole Polytechnique, INRIA, Universite Paris-Sud.

Boden, M.A., 1979. The computational metaphor in psychology, in: Bolton, N. (Ed.), Philosophical Problems in Psychology. Methuen.

Bohus, D., Horvitz, E., 2010. Computational Models for Multiparty Turn Taking. Technical Report. Microsoft.

Bohus, D., Horvitz, E., 2011. Multiparty turn taking in situated dialog: Study, lessons, and directions, in: Proceedings SigDial2011.

Bos, J., Klein, E., Lemon, O., Oka, T., 2003. Dipper: Description and formalisation of an information-state update dialogue system architecture, in: 4th SIGdial Workshop on Discourse and Dialogue, pp. 115–124.

Louis ten Bosch, Nelleke Oostdijk, L.B., 2005. On temporal aspects of turn taking in conversational dialogues. Speech Communication 47(12).

Bradley, M.M., Lang, P.J., 1994. Measuring emotion: the self-assessment manikin and the semantic differential. Journal of Experimental Psychiatry & Behavior Therapy 25, 49–59.

Brown, P., Levinson, S., 1987. Politeness - Some universals in language usage. Cambridge University Press.

Burghouts, G., Akker, R.o.d., Heylen, D., Poel, M., Nijholt, A., 2003. An action selection architecture for an emotional agent, in: Russell, I., Haller, S. (Eds.), Recent Advances in Artificial Intelligence, AAAI Press. p. 293297.

Cassell, J., Thórisson, K.R., 1999. The power of a nod and a glance: Envelope vs. emotional feedback in animated conversational agents. International Journal of Applied Artificial Intelligence 13(4-5), 519–538.

Castelfranchi, C., Falcone, R., 2003. From automaticity to autonomy: the frontier of artificial agents, in: Hexmore, H., Castelfranchi, C., Falcone, R. (Eds.), AgentAutonomy. Kluwer Academic Publishers, pp. 104–125.

Clark, H., 1996. Using Language. Cambridge University Press.

Clark, H.H., Krych, M.A., 2004. Speaking while monitoring addressees for understanding. Journal of Memory and Language 50(1), 62–81.

Cowie, R., Sussman, N., Ben-Zeev, A., 2011. Emotion: Concepts and definitions, in: Petta, P., Pelachaud, C., Cowie, R. (Eds.), Emotion-oriented systems: the Humaine handbook. Springer Verlag, pp. 9–30.

Cowley, S.J., 1998. Of timing, turn-taking, and conversations. Journal of Psycholingitistic Research 27(5), 541–571.

Dalgleish, T., 2004. The emotional brain. Nature 5, 582–589.

Duncan, S., 1972a. Some signals and rules for taking speaking turns in conversations. Journal of Personality and Social Psychology 23, 283–92.

Duncan, S., 1972b. Some signals and rules for taking speaking turns in conversations. Journal of personality and social psychology , 283–292.

Edelsky, C., 1981. Who's got the floor? Language and Society 10(3), 383–421.

Frijda, N.H., 1986. The emotions. Cambridge University Press, Cambridge, UK.

Gebhard, P., Klesen, M., Rist, T., 2003. Adding the emotional dimension to scripting character dialogues, in: Proc. of the 4th International Working Conference on Intelligent Virtual Agents (IVA'03), Kloster Irsee, pp. 48–56.

Goffman, E., 1967. Interaction Ritual: Essays on Face-to-Face Behavior. Random House.

Goldberg, J., 1990. Interrupting the discourse on interruptions An analysis in terms of relationally neutral, power- and rapport-oriented acts. Journal of Pragmatics 14, 883–903.

Goldie, P., 2000. The emotions: a philosopical exploration. Clarendon Press.

Grice, J.P., 1975. Logic and conversation. Syntax and Semantics 3: speech acts (P. Cole and J.L.Morgan (eds) , 41–58.

Harel, D., 1987. Statecharts: A visual formalism for complex systems. Science of Computer Programming 8, 231–274.

Hayashi, R., 1991. Floor structure of english and japanese conversation. Journal of Pragmatics 16, 1–30.

Heylen, D., op den Akker, H., ter Maat, M., Petta, P., Rank, S., Reidsma, D., Zwiers, J., 2011a. On the nature of engineering social artificial companions. Applied Artificial Intelligence 25, 549–574.

Heylen, D., Bevacqua, E., Pelachaud, C., Poggi, I., Gratch, J., Schröder, M., 2011b. Generating Listening Behaviour, in: Cowie, R., Pelachaud, C., Petta, P. (Eds.), Emotion-Oriented Systems. Springer Berlin Heidelberg. Cognitive Technologies, pp. 321–347.

Hofs, D., Theune, M., op den, R.A., 2010. Natural interaction with a virtual guide in a virtual environment: A multimodal dialogue system. Journal on Multimodal User Interfaces 3, 141–153. Open Access.

Hollnagel, E., 2003. Is affective computing an oxymoron? Int. J. Hum.-Comput. Stud. 59, 65–70.

Kilpatrick, P., 1986. Turn and Control in Puerto Rican Spanish Conversation. Technical Report. University of Puerto Rico.

Kronlid, F., 2006. Turn taking for artificial conversational agents, in: Cooperative Information Agents X. Springer Berlin / Heidelberg. volume 4149 of *Lecture Notes in Computer Science*, pp. 81–95.

Kronlid, F., 2008. Steps towards Multi-Party Dialogue Management. Ph.D. thesis. The Graduate School of Language Technology, University of Gotenburg.

Kronlid, F., Lager, T., 2007. Implementing the information-state update approach to dialogue management in a slightly extended scxml, in: Artstein, R., Vieu, L. (Eds.), Proceedings of the 11th International Workshop on the Semantics and Pragmatics of Dialogue (DECALOG), pp. 99–106.

Lang, P.J., 1980. Behavioral treatment and bio-behavioral assessment: computer applications, in: Sidowski, J., Johnson, J., Williams, T. (Eds.), Technology in mental health care delivery systems. Norwood, NJ: Ablex, pp. 119–l37.

LeDoux, J.E., 1996. The Emotional Brain: the Mysterious Underpinning of Emotional Life. Simon & Schuster.

Lerner, G.H., 2003. Selecting next speaker: The context-sensitive operation of a context-free organization. Language in Society 32, 177–201.

ter Maat, M., 2011. Responce Selection and Turn-taking for a Sensitive Artificial Listening Agent. Ph.D. thesis. University of Twente.

ter Maat, M., Heylen, D., 2009. Turn management or impression management?, in: Ruttkay, Z., Kipp, M., Nijholt, A., Vilhjalmsson, H. (Eds.), Intelligent virtual agents, Springer-Verlag. pp. 467–473.

Marinetti, C., Moore, P., Lucas, P., , Parkinson, B., 2011. Emotions in social interactions: Unfolding emotional experience, in: Petta, P., Pelachaud, C., Cowie, R. (Eds.), Emotion-oriented systems: the Humaine handbook. Springer Verlag, pp. 31–46.

Marsella, S., Gratch, J., 2009. EMA: A process model of appraisal dynamics. Cognitive Systems Research 10, 70–90.

Novick, D.G., Hansen, B., Ward, K., 1996. Coordinating turn-taking with gaze, in: Proceedings of ICSLP-96.

O'Connell, D.C., Kowal, S., Kaltenbacher, E., 1990. Turn-taking: A critical analysis of the research tradition. Journal of Psycholinguistic Research 19, No. 6, 345–373.

Ortony, A., Clore, G.L., Collins, A., 1990. The Cognitive Structure of Emotions. Cambridge University Press, Cambridge.

Peters, C., Pelachaud, C., Mancini, M., Bevacqua, E., 2005. Engagement capabilities for ecas, in: Proc. Workshop "Creating bonds with ECAs", AAMAS'2005.

Petta, P., Pelachaud, C., (eds.), R.C., 2011. Emotion-oriented systems: the Humaine handbook. Springer Verlag.

Picard, R.W., 2003. What does it mean for a computer to “have” emotions?, in: Trappl, R., Petta, P., Payr, S. (Eds.), Emotions in Humans and Artifacts. MIT Press.

Picard, R.W., Klein, J., 2002. Computers that recognise and respond to user emotion: theoretical and practical implications. Interacting with Computers 14, 141–169.

Poggi, I., Pelachaud, C., de Rosis, F., Carofiglio, V., de Carolis, B., 2005. Greta: a believable embodied conversational agent, in: Paggio, P., Jongejan, B. (Eds.), Multimodal Communication in Virtual Environments, pp. 27–45.

Prendinger, H., Ishizuka, M., 2001. Social role awareness in animated agents, in: Proceedings of the fifth international conference on Autonomous agents, pp. 270–277.

Prendinger, H., Ishizuka, M., 2002. Scream: scripting emotion-based agent minds, in: Proceedings of the first international joint conference on Autonomous agents and multiagent systems: part 1, ACM, New York, NY, USA. pp. 350–351.

Raux, A., Eskenazi, M., 2007. A multi-layer architecture for semi-synchronous event-driven dialogue management. Automatic Speech Recognition & Understanding , 514–519.

Raux, A., Eskenazi, M., 2009. A finite-state turn-taking model for spoken dialog systems, in: Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the ACL, ACL, Stroudsburg, PA, USA. pp. 629–637.

Reidsma, D., de Kok, I., Neiberg, D., Pammi, S., van Straalen, B., Truong, K., van Welbergen, H., 2011. Continuous interaction with a virtual human. Journal on Multimodal User Interfaces 4, 97–118.

Reidsma, D., Truong, K.P., van Welbergen, H., Neiberg, D., Pammi, S., de Kok, I.A., 2010. Continuous interaction with a virtual human, in: Proceedings of the 6th International Summer Workshop on Multimodal Interfaces (eNTERFACE’10), University of Amsterdam, Amsterdam. pp. 24–39.

Rist, T., Baldes, S., Gebhard, P., Kipp, M., Klesen, M., Rist, P., Schmitt, M., 2002. Crosstalk: An interactive installation with animted presentation agents, in: Proceedings of 2nd Conference on Computational Semiotics for Games and New Media, Universitt Augsburg, pp. 61–67.

Robinson, L.F., Reis, H.T., 1989. The effects of interruption, gender, and status on interpersonal perceptions. Journal of Nonverbal Behavior 13, 141–153.

Sacks, H., Schegloff, E., Jefferson, G., 1974. A simplest systematics for the organization of turn-taking for conversation. Language 50, 696–735.

Schegloff, E.A., 1968. Sequencing in conversational openings. American Anthropologist 70, 1075–1095.

Schegloff, E.A., 2000. Overlapping talk and the organization of turn-taking for conversation. Language in Society 29, 1–63. http://journals.cambridge.org/article-S0047404500001019.

Shearer, J., Olivier, P., Boni, M.D., Hurling, R., 2007. Exploring persuasive potential of embodied conversational agents utilizing *ynthetic* embodied conversational agents, in: PERSUASIVE, pp. 210–213.

Simon, H.A., 1967. Motivational and emotional controls of cognition. Psychological Review 74, 29–39.

Sloman, A., 2001. Beyond shallow models of emotion. Cognitive Processing 2(1), 177198.

Sloman, A., Chrisley, R., Scheutz, M., 2003. Architectural basis of affective states and processes, in: Fellous, Arbib (Eds.), Who Needs Emotions?: The Brain Meets the Machine. Oxford University Press.

Steunebrink, B., Vergunst, N., Mol, C.P., Dignum, F., Dastani, M., Meyer, J.J., 2008. A generic architecture for a companion robot, in: Filipe, J., Cetto, J., Ferrier, J. (Eds.), Proc. 5th Int. Conf. on Informatics in Control, Automation and Robotics (ICINCO'08), pp. 315–321.

Stivers, T., Enfield, N., Brown, P., Englert, C., Hayashi, M., Heinemann, T., Hoymann, G., Rossano, F., de Ruiter, J.P., Yoon, K.E., Levinson, S.C., 2009. Universals and cultural variation in turn-taking in conversation. PNAS 106(26), 10587–10592.

ter Maat, M., Heylen, D.K.J., 2011. Flipper: An information state component for spoken dialogue systems, in: Vilhjálmsson, H., Kopp, S., Marsella, S., Thórisson, K. (Eds.), Proceedings of the 10th international conference on Intelligent Virtual Agents (IVA 2011), Reykjavik, Iceland, Springer Verlag, Berlin. pp. 470–472.

Theune, M., Rensen, S., op den, R.A., Heylen, D., Nijholt, A., 2004. Emotional characters for automatic plot creation, in: Göbel, S., Spierling, U., Hoffmann, A., Iurgel, I., Schneider, O., Dechau, J., Feix, A. (Eds.), Technologies for Interactive Digital Storytelling and Entertainment, Springer Verlag, Berlin. pp. 95–100. Imported from HMI.

Thórisson, K., Gislason, O., Jonsdottir, G., Thorisson, H., 2010. A multiparty multimodal architecture for realtime turntaking, in: Allbeck, J., Badler, N., Bickmore, T., Pelachaud, C., Safonova, A. (Eds.), Intelligent Virtual Agents. Springer Berlin / Heidelberg. volume 6356 of *Lecture Notes in Computer Science*, pp. 350–356.

Thorisson, K.R., 2002. Natural turn-taking needs no manual: Computational theory and model, from perception to action, in: Multimodality in Language and Speech Systems. Kluwer Academic Publishers, p. 173207.

Traum, D., Marsella, S.C., Gratch, J., Lee, J., Hartholt, A., 2008. Multiparty, multi-issue, multi-strategy negotiation for multi-modal virtual agents, in: Proceedings of the 8th international conference on Intelligent Virtual Agents, Springer-Verlag, Berlin, Heidelberg. pp. 117–130.

Traum, D., Rickel, J., 2002. Embodied agents for multiparty dialogue in immersive virtual worlds, in: Proceedings of the first international joint conference on Autonomous agents and multiagent systems: part 2., ACM Press.

Traum, D., Swartout, W., Marsella, S., Gratch, J., 2005. Fight, flight, or negotiate: Believable strategies for conversing under crisis, in: Intelligent Virtual Agents, Springer. pp. 52–64.

Vilkki, L., 2006. Politeness, face and facework: Current issues, in: Suominen, M. (Ed.), A Man of Measure: Festschrift in Honour of Fred Karlsson on his 60th birthday. Linguistic Association of Finland, p. 322332.

Wang, Z., Lee, J., Marsella, S., 2011. Towards more comprehensive listening behavior: beyond the bobble head, in: Proceedings of the 10th international conference on Intelligent virtual agents, Springer-Verlag, Berlin, Heidelberg. pp. 216–227.

Wieland, M., 1991. Turn-taking structure as a source of misunderstanding in french-american cross-cultural conversation. Pragmatics and Language Learning 2, 101–118.

Wilks, Y. (Ed.), 2009. Close Engagements with Artificial Companions: Key social, psychological, ethical and design issues. John Benjamins Publ. Cy.

Yannakakis, G.N., Togelius, J., Khaled, R., Jhala, A., Karpouzis, K., Paiva, A., Vasalou, A., 2010. Siren: Towards adaptive serious games for teaching conflict resolution, in: Proceedings 4th European Conference on Games Based Learning ECGBL10.

Yngve, V., 1970. On getting a word in edgewise, in: Papers from the sixth regional meeting of the Chicago Linguistic Society, Chicago: Chicago Linguistic Society. pp. 567–77.