Response Selection and Turn-taking for a Sensitive Artificial Listening Agent

Mark ter Maat

PhD dissertation committee:

Chairman and Secretary:
  Prof. dr. A.J. Mouthaan, University of Twente, NL
Promotor:
  Prof. dr. A. Nijholt, University of Twente, NL
Assistant-promotor:
  Prof. dr. D.K.J. Heylen, University of Twente, NL
Members:
  Prof. dr. C. Pelachaud, CNRS LTCI, TELECOM Paris Tech, FR
  Prof. dr. D. Schlangen, University of Bielefeld, DE
  Prof. dr. E.J. Krahmer, University of Tilburg, NL
  Prof. dr. V. Evers, University of Twente, NL
  Prof. dr. F.M.G. de Jong, University of Twente, NL
  Dr. J. Zwiers, University of Twente, NL
Paranymphs:
  Bart van Straalen
  Arjan Gelderblom

Human Media Interaction group
The research reported in this dissertation has been carried out at the Human Media Interaction group of the University of Twente.

CTIT Dissertation Series No. 11-211
Center for Telematics and Information Technology (CTIT)
P.O. Box 217, 7500 AE Enschede, The Netherlands. ISSN: 1381-3617.

SEMAINE
The research reported in this thesis has been carried out in the SEMAINE (Sustained Emotionally coloured Machine-human Interaction using Nonverbal Expression) project. SEMAINE is sponsored by the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 211486.

SIKS Dissertation Series No. 2011-48
The research reported in this thesis has been carried out under the auspices of SIKS, the Dutch Research School for Information and Knowledge Systems.

ISBN: 978-94-6191-105-6

Response Selection and Turn-taking for a Sensitive Artificial Listening Agent

DISSERTATION

to obtain
the degree of doctor at the University of Twente,
on the authority of the rector magnificus,
prof. dr. H. Brinksma,
on account of the decision of the graduation committee
to be publicly defended
on Wednesday, November 30, 2011 at 14:45

by

Mark ter Maat

born on December 20, 1983

in Ede, The Netherlands

This thesis has been approved by:
Prof. dr. ir. Anton Nijholt, University of Twente, NL (promotor)
Prof. dr. Dirk Heylen, University of Twente, NL (assistant-promotor)

ISBN: 978-94-6191-105-6

Acknowledgements

Vier jaar. Soms lijkt het een eeuwigheid, soms vliegt het voorbij, en dit heeft vooral te maken met waar je mee bezig bent (is die deadline nu alweer?). Gelukkig mocht ik werken aan een geweldig project in een geweldige vakgroep, waardoor de tijd (helaas) veel te snel ging. Het was af en toe wel moeilijk balanceren tussen werken aan Sal en 'echt' onderzoek doen, maar dat laatste is gelukt met als resultaat dit boekje.

Als eerste moet ik hier toch Dirk voor bedanken. Hij heeft me de vrijheid gegeven om mijn eigen dingen te doen, en me ondertussen toch een goede richting in te sturen. Ook kon ik altijd terecht als ik met een conceptueel probleem zat, ik niet kon kiezen wat de beste methode was, of om te vragen waar de beste theewinkels in Parijs te vinden zijn. Dirk, bedankt.

Switching to English for a while, I want to thank Anton, for being my promotor and for forcing me each year to think about my goals of next year. I want to thank the members of my committee as well, for taking the time to read and comment on this thesis which has an astoundingly little amount of pictures.

Next, I want to thank the members of the project I have been working on: SEMAINE. Marc, you were a great project leader, and I learned a lot from you about working on big projects. Also, I especially want to thank the other 'junior members' of the team: Michel, Hatice, Elisabetta, Florian, Martin and Sathish. Our developer-meetings were not only really productive, but also a lot of fun, and I will never forget our excessive dinner in Paris.

Uiteraard wil ik ook al mijn collega's bij HMI bedanken, maar een aantal in het bijzonder. Bart, wij zijn ongeveer gelijk begonnen met een onderwerp dat heel erg op elkaar lijkt, maar ook heel verschillend is. Dit heeft er voor gezorgd dat we heel wat uurtjes hebben zitten bakkeleien over beliefs over beliefs en andere conceptuele dingen, en was je de ideale betatester. En ook hebben we heel wat gezellige en ontspannende avonden gehad met z'n drien. Bart, bedankt.

Een aantal artikelen heb ik samen met Khiet gemaakt, en haar neigingen om alle gaatjes perfectionistisch af te dichten heeft mij ook een heleboel geleerd. Dennis, jij hebt mij vooral geleerd om op een hoog niveau verder te denken, maar ik kon ook altijd binnenwandelen om een leuke demo te laten zien. Ronald, behalve comic relief bij HMI en onze 'vakantie' naar philadelphia kon ik ook altijd terecht voor een praktisch gesprek en voor tips over classifiers en experiment-opzetten. Iwan, ook wij hebben leuke tripjes gemaakt en hele leuke gesprekken gehad, en ik blijf erbij dat we onze anekdote-agent nog steeds een keer moeten laten maken. En Danny, bedankt voor het zijn van een geweldige kamergenote (als je er was), voor het uitwisselen van willekeurige nieuwsberichtjes, of om een keer te praten over mijn onderzoek met iemand uit een hele andere hoek.

Uiteraard wil ik ook Lynn bedanken, voor het doorlezen en verbeteren van alle kleine puntjes in mijn proefschrift. Hopelijk is mijn dt-probleem nu opgelost. Charlotte en Alice, jullie waren er altijd voor het beantwoorden van organisatorische vragen, zelfs de onbenullige. En Hendri, ik heb ook regelmatig voor jouw deur gestaan met weer een vraag over hardware of software. Allemaal heel erg bedankt.

En dan muziek. Zoals ook dit boekje duidelijk laat zien kom ik altijd weer uit op muziek, en mijn grootste uitlaatklep daarvan is Animato. En dit komt ook heel erg door de mensen daar. Annemieke, Arjen, Cor, Ellen, Lisette, Louise, Marieke, Milenne, en Raymond, jullie maken alle repetities nog leuker dan ze al zijn, en zorgen er samen met onze reguliere feestjes voor dat ik vaak ook met iets anders bezig kan zijn dan mijn onderzoek.

Ik wil natuurlijk ook mijn paranimfen bedanken, Bart en Arjan. Ik ben vereerd dat jullie naast me willen staan op het podium. Arjan, ook met jou heb ik af en toe heel goed kunnen praten over mijn onderzoek, of computer-stuff in het algemeen, en ook hebben wij heel wat gezellige avonden (met of zonder Palm) gehad.

Verder wil ik mijn familie bedanken, met name mijn ouders en broertje en zusje, voor jullie steun, zinnige en onzinnige gesprekken als ik weer eens thuis was.

En als laatste, maar ook als belangrijkste, wil ik Saskia bedanken. Sas, bedankt voor je liefde, je steun, de cover, je talent om mijn aangeboren aanleg voor chaos wat gestructureerder te krijgen, en gewoon voor een geweldige tijd met je samen. Ik was niet altijd makkelijk, als ik wel thuis was maar mijn hoofd nog halverwege een paper of een programmeer- of schrijfprobleem, maar jij wist me altijd weer terug te trekken. Heel erg bedankt, ik ben blij dat je mij gevonden hebt.

Contents

Part I

Sensitive Artificial Listeners

1 Virtual characters

1.1 Interacting with a computer

Humans have a very peculiar relationship with machines, and especially with computers. Designed to make our lives easier, and continuously changing in how we interact with them. From command-line interfaces to mouse-based interfaces, but we still have to learn how to interact with them. First we had to memorize commands, but nowadays, with the mouse and graphical interfaces we have to learn how to split our intention into smaller, clickable elements: which elements these are differs per task. When facing a new task the user has to find out how to translate his or her intentions into actions that the system understands. Because of this, researchers and designers are focussing more and more on making interaction with a computer more natural. Even though communication with a machine is inherently not ‘natural’, we prefer to interact with them without learning new skills. This means we have to use types of communication that we already know. An obvious choice would be how we communicate with other persons. We are social beings, and we learn to communicate with other human beings from the day we are born, so why can we not use this method of communication to make our intentions clear to a machine?

The most simple answer to this question is because it is really hard for a machine to understand language, and it takes time to understand human behaviour well enough to teach a computer to understand human intentions. Human speech is highly ambiguous, and intimately tied to the context. And non-verbal signals that are sent along with the speech may also change the meaning in several ways. Also when humans communicate, misunderstandings arise regularly, which humans usually repair swiftly. Computers have more difficulty with this.

Nevertheless, researchers have developed all kinds of dialogue systems, and usually simplified matters a lot to make them workable. For example, Matheson et al. (2000) describe a dialogue system that only uses the speech modality, and from the speech only uses the detected words. The authors use the dialogue move engine toolkit TrindiKit (Traum et al. (1999)), which can be used for dialogue systems that perform a certain task for a user. It uses natural language to ask questions in order to get the required information, and when the user answers the necessary information is extracted. As an example, the authors provide a sample dialogue that should be possible with their dialogue system. In this example, the user wants a route from the dialogue system. The system then asks for all the information it needs: the start location, the end location, departure time, and whether the user wants the quickest or the shortest route. Also, when the system is not sure about an answer, it asks for a clarification. For example, when the user specifies the location 'Edwinstowe', the system asks 'Edwinstowe in Nottingham?'.

Dialogue systems such as this make it possible to use natural language to interact with a computer, but there are still some limitations. First of all, this particular system does basically nothing more than 'slot-filling': it needs to fill in certain fields (slots), and asks the user questions about each field until all fields are filled in. This way of proceeding in a conversation is only 'natural' in very specific situations (for example, when buying a train ticket) in which the computer needs specific information from the user. Secondly, a lot of natural communication mechanisms are missing. For example, the dialogue as presented by Matheson et al. is a sort of ping-pong, with a turn for the user, a turn for the machine, etcetera. But what happens when the user interrupts the system? And if the user is in the middle of a sentence and a misunderstanding arises, why not interrupt the user quickly to ask for a clarification about the first part, and then let him continue? Also, the system could improve the communication by occasionally signalling to the user that it is still following and understanding him by giving a backchannel signal. Thirdly, even with speech only, a system can detect and use a lot more than only the words (and with that the communicative intentions) from the user. By detecting prosodic information and non-linguistic items (for example backchannels, such as "uhuh"), a system could detect which words are emphasized (and therefore more important), what the current emotional state of the user is, whether the user shows understanding, and so on. Finally, the system in the example only uses one modality (speech). By adding more modalities, for example by adding a camera and a virtual character, the range of interaction possibilities increases enormously. Suddenly, the user and the machine can look each other in the eyes, gesture to each other, and show facial expressions. Detecting multi-modal behaviour also increases the likelihood of detecting the correct intention from the user.

Of course, using more modalities than only the content of the user's speech brings its own set of problems and challenges. If the virtual agent becomes too humanlike, people will expect the character to communicate as a human person too. This means that everything the virtual character does influences how a user perceives that character, and if the character makes a 'communicative error' the quality of the conversation goes downhill fast. A communicative error can be almost anything: the wrong facial expression which changes the meaning of an utterance, starting to speak at the wrong time, failing to give feedback or giving it too late, and saying something totally unrelated and inappropriate as a reaction to what the user just said.

The context of this thesis is the development of a virtual character, or to be more specific, a virtual listening agent. The complete system is called Sal, for Sensitive Artificial Listener, and consists of four different characters: Poppy, the optimistic character, Obadiah, the depressed character, Spike, the aggressive character, and Prudence, the rational character. These characters try to motivate the user to speak for as long as possible, and to get the user into the same emotional state as they themselves are in. Occasionally, the listening agent should of course respond (you cannot expect the user to keep talking for eternity to a virtual head that only nods and says 'yeah'). This response should focus on motivating the user to continue speaking, either by giving feedback on what was just said ('That sounds great'), asking for more information ('When will that happen?'), or changing the subject ('Do you have any other plans?'). The contribution of this thesis is the turn-taking behaviour of the different characters, and the selection of an appropriate response that is mainly based on the non-verbal behaviour of the user.

1.2 This thesis

In this section, we will describe the two topics of this thesis, namely the perception of turn-taking, and response selection.

1.2.1 Turn-taking

Turn-taking behaviour in virtual agents is often a problematic aspect. When humans interact with each other, turn switches seem fluent and without any problems (although in reality humans are just good at fixing problems swiftly). Pauses between turns are often short, and overlapping speech rarely causes a lot of problems. However, when trying to implement this fluent turn-taking behaviour in virtual agents it gets clear that it is not as straightforward as it seems. Agents have to be capable of detecting when users want the turn, when they start a turn, and when they finish a turn or almost do so, and this is not always easy.

When looking at the turn-taking behaviour of virtual agents, it is remarkable that a lot of systems do not describe this conversational aspect in detail. Sometimes this is because the implemented modalities do not really support turn-taking. For example, in the systems of Schulman and Bickmore (2009) and Bosch et al. (2009) users have to click on the button that matches what they want to say to the agent. Naturally, turn-taking is not an issue here anymore. And in the museum guide described by Satoh (2008), users have to communicate with the agent by moving to a certain location (or moving away to stop the conversation).

Several virtual agent systems use text as input, which also makes the turn-taking aspect less interesting — when the user presses Enter or clicks on the 'send' button, the text is sent to the system and is treated as a finished turn. Max (Kopp et al., 2005) goes one step further: the system continuously monitors the keyboard, and typing is considered the same as speaking. When Max is talking and the user starts typing, Max immediately stops speaking.

Only when speech is added as an input modality, does turn-taking really become an issue. Even determining when the user finishes the turn or offers it to the agent is not as easy as humans sometimes make it seem. Only relying on pauses does not always work, since people often pause within their own turn too (Edlund et al. (2005)). This means that the agent should either use the words of the user's sentence or the detected prosody to determine whether a turn is finished or not. De Ruiter et al. (2006)

demonstrate that people can still recognize the end of a turn when prosodic information is removed, but perform a lot worse when the words were removed and only the prosodic information was audible. However, this means relying on (often inaccurate) speech recognition and difficult natural language understanding. Therefore, usually only prosodic information is used with end of turn detection, for example by Jonsdottir et al. (2008). Unfortunately, very few papers about virtual agents with speech input explain in detail how end of turn detection was implemented.

However, what *is* sometimes described is what the agents do when they detect that the user is speaking while the agents are talking too. Remarkably, a lot of these agents — for example Max (Kopp et al., 2005), the Virtual Guide from Theune et al. (2007), and Rea (Cassell et al., 1999) — respond extremely politely by immediately stopping their speech, no matter what they were saying and what the user is doing. An erroneous detection of speech — for example, a little cough — could completely stop an explanation, and stopping two or three words before the end of an utterance hardly seems useful.

This brings us to one of the most important points about turn-taking in this thesis: what about the user's perception of the agent? The developers of the virtual agents mentioned above make the agent stop speaking when it detects that both the agent and the user are talking at the same time, with the intention to make it more polite. But in doing so, they could change a lot more than just the perception of politeness. For example, the agent could be seen as passive and weak. And when stopping its utterance when it was not necessary, the agent could be perceived as shy or indecisive. The question is whether this is desirable.

The first part of this thesis is about the perception of different turn-taking strategies. We look at strategies for starting a turn — which can take place just before the other person is finished, exactly at that moment, or after a small period of silence — and strategies for how to behave when overlapping speech is detected — which can be to stop speaking, to continue normally as if nothing has happened, or to continue with a raised voice to stop the other person from speaking. We describe several experiments in which we tried to find out how these strategies influence the perception of the user on the receiving end.

1.2.2 Response selection

No matter how simplistic or complex a dialogue system is, it needs a method of determining what to say during interactions with the user. In a small domain, it is usually sufficient to create a Finite State Machine (FSM), in which all possible interactions are predefined. For example, for simple receptionists, an FSM is sufficient for most of its tasks, as shown in the virtual receptionists Mack (Cassell et al. (2002)) and Marve (Babu et al. (2006, 2005)). These agents have fairly straightforward tasks, such as giving directions, answering questions about the environment, and delivering messages to people interacting with it. This is exactly the kind of interaction that is perfect for state machines: short topics, with clearly different states that the conversation can be in, and simple transitions between these conversation states.

A similar method is writing out the complete dialogue tree, including all possible deviations of the dialogue. This is especially useful if there only are a limited number of user dialogue moves. For example, in the dialogue system described by Schulman and Bickmore (2009), during the interaction the user has to select a response from a small list. This makes it possible to create the complete dialogue tree beforehand. And in the museum guide from Satoh (2008) the user has to respond by standing on a certain spot left or right of the agent. What this method boils down to is that the agent tells a story and at some points offers the user the option of influencing the direction of the story. And sometimes it is not even that, but the agent only responds to the choice of the user with one or two sentences, and then continues its story.

However, a downside of these methods is the lack of flexibility and extensibility. For example, the complete dialogue needs to be written out beforehand, which is only feasible in small and very specific domains. Also, conversations could feel unnatural, because everything is scripted and deterministic.

A method with more interaction is called *slot filling*, for example demonstrated by Nijholt and Hulstijn (2000). With this method, the agent offers some kind of service — for example giving directions — but in order to do this it needs information, in the form of information fields, called slots, that need to be filled in. To fill these slots the agent can ask questions of the user, either to get the required information or to verify information that it already has, but is not sure of. This method is most useful for service agents that need information from the user in order to provide a certain service, for example giving directions, selling train tickets, etcetera.

A more general purpose method is used by Traum et al. (2007) and McCauley and D'Mello (2006). Instead of specifying the complete dialogue and all transitions, the system is fed all responses it can give to a user, and for each response a set of user utterances that could lead to that response. When the agent receives input from the user, it tries to find the stored user utterance that is statistically the most similar, and then returns the corresponding response. This method is especially powerful for question-answering systems, in which the agent does not need to keep track of the flow of the conversation, but only has to answer the questions of the user.

If the developers want more control of the agent's behaviour, using hand-crafted templates might be a good alternative (see for example the work of Kopp et al. (2005)). Hand crafting templates is manual labour and seems to suffer from the same problem as state machines and dialogue trees, that it is a lot of work for a large domain. But in contrast with these methods, it is not necessary to specify all possible transitions when using templates, which makes adding new behaviour relatively easy since the previous models do not need to be modified much.

In short, the most efficient and effective method of response selection depends on the domain and the context. It depends on the number of topics, the modalities of the user input, what is known beforehand about the dialogue structure, etcetera. But what to do when the context and domain is unknown beforehand, or even worse, unknown even during interaction? Simply said, is it possible to have an interaction with a person without understanding him or her? Is it enough to have information about the prosody, head movements and facial expressions to give appropriate and relevant responses?

In the second part of this thesis, we discuss this in the context of the SEMAINE project, in which we built a virtual listening agent, capable of giving appropriate responses without knowing anything about the actual content. We explain different methods we tried to craft rules that select responses based on the user's input, and after that we explain how we used Wizard of Oz data to make the agent learn rules.

2 SEMAINE

The research that is described in this thesis was carried out in the SEMAINE project. In order to better understand this research, why it was carried out this way, and its consequences, the SEMAINE-context should be clear. This section explains the SEMAINE system: why we want sensitive artificial listeners, the architecture of Sal and its components, and a brief overview of the evaluation results.

2.1 Sensitive Artificial Listening agents

The concept of a Sensitive Artificial Listening agent (a SAL) was first proposed by Balomenos et al. (2003) in the context of the ERMIS project. They explained that SAL is a descendant of ELIZA (Weizenbaum, 1966), in the sense that it does not really understand the user but uses tricks to pretend understanding. SAL analyses the user's voice and facial expressions, extracts signs of emotion and uses these to select a stock response. The aim of Sal is to keep the user motivated to keep on talking.

ERMIS stands for Emotion-Rich Man-machine Interaction Systems, and aims to systematically analyse speech and facial input signals, in order to "extract parameters and features which can then provide MMI (Man-Machine Interaction) systems with the ability to recognise the basic emotional states of their users and respond to them in a more natural and user friendly way". A SAL was built to serve as a testbed for the techniques developed in the project. But besides being a testbed, the authors also noted that such a system has other uses too. Like ELIZA, such systems are fun, and perhaps mildly therapeutic. And a more serious use is that they provide an environment in which emotions can occur in a natural conversation setting.

In the HUMAINE Network of Excellence[1], this approach was also used extensively. HUMAINE aimed to get a better understanding of emotions and issues that are involved with using emotions with virtual characters. It addressed issues such as the theory of emotional processes, automatic detection and synthesis of emotions, how emotion affects cognition, and the gathering of emotional data (Schröder and Cowie, 2005). For this last issue, part of the emotional data that was collected was created

[1] http://emotion-research.net/

with the SAL scenario (McKeown et al., 2010). They created four different characters, each with a different 'personality' and a corresponding set of responses. These four characters are:

- **Poppy** – Optimistic and outgoing
- **Prudence** – Pragmatic and practical
- **Obadiah** – Depressed and gloomy
- **Spike** – Confrontational and argumentative

Besides motivating the user to continue speaking, these characters have the additional goal to draw the user to their emotional state. Thus, Poppy is constantly trying to make the user happy and optimistic, while Spike tries to get the user in an angry state of mind. These characters are played in a Wizard of Oz (WOz) setting, in which the user talks with a computer system, which is controlled by a human (the wizard) behind the scenes. This wizard has a set of scripts, based on the current character and the emotional state of the user: positive-active, positive-passive, negative-active or negative-passive. The wizard also has some scripts for different states of the conversation, such as the start of the conversation. During the conversations, the wizard watches and listens to the user and, based on the current script, determines which sentence of that script the system should say. Even with a fixed set of possible responses, this setup turned out to be capable of maintaining a conversation, sometimes for even up to half an hour.

In the HUMAINE project (Schröder et al., 2008), the SAL setup was used as a means to gather emotional data, and the agent was controlled by a human. However, the succeeding project *SEMAINE* aimed to build a fully automatic version of a SAL: a multi-modal dialogue system which focusses on non-verbal skills. It contains the same four characters as in HUMAINE, and aims to automatically detect emotional features from the voice and the face and respond to these, while also showing understanding by giving backchannels.

So, in total, at least three projects have worked with the SAL concept. One might wonder if it is possible to sustain a conversation with a virtual agent that does not really understand what the conversation is about. But humans can do it too, so why not virtual agents? It is possible for two humans to have a conversation, while one participant pays little or no attention to what the other person is really saying. This is demonstrated by ELIZA (Weizenbaum, 1966), which can keep a conversation going by incorporating the user's input into its output with a set of generic patterns. For example, the sentence 'I like to X', in which X can be anything, could lead to the response 'Tell me, why do you like to X?'. In most situations, this suggests that ELIZA understands the user. The SAL concept uses emotions instead of words or phrases to determine what to say. And McKeown et al. (2010) show that with these limited set of responses it is possible to sustain a conversation.

A second issue could be: why would we want this? A first reason was already mentioned earlier in this section, namely that it serves as a setting that allows natural conversations which could induce emotions from the user. A second, more technical reason is that it serves as a testbed and a demonstrator of techniques that have been developed that deal with emotions and affect. These techniques could later be applied in other virtual agents, for example to provide them with the possibility to detect emotions. But even the SAL concept itself could be useful for other virtual characters. It gives us the opportunity to learn how agents should react to detected emotions. And finally, it can serve as some kind of fallback mechanism when the actual content of the user's sentence is not understood or out of the agent's scope. When this happens, it can fall back on responses that are still appropriate to the user's emotional state, even though the actual content is unknown.

2.2 The SEMAINE virtual agent

The SEMAINE virtual agent uses the SAL concept: its main focus is listening behaviour and motivating the user to speak for as long as possible. Additionally, there are four versions of the agent, and each version has a different character with a different emotional state. The characters have been described in the previous section, and from now on, we will refer to the complete system as *Sal*. In this section we explain how Sal works, which components it consists of and how they communicate with each other.

2.2.1 Global architecture

Sal consists of six main components, which are all shown in Figure 2.1: audio input, video input, dialogue management, the virtual agent, the audio synthesis, and the API that connects all modules together (the arrows).

The audio and video input function as the eyes and ears of Sal; they capture what the user says and does. The audio input uses a microphone to capture every sound the user makes: speaking, laughing, coughing, etcetera. With this data, it extracts low-level features such as the F0-frequency and the energy, affective features such as valence, arousal and interest, and non-verbal features such as laughs and sighs. The video input component records the user from the shoulders up, and extracts features such as head position, rotation, movement, gestures (nods and shakes), and facial expressions.

These extracted features are sent to the dialogue management component. The first task of this component is to interpret the detected behaviour given the current context. For example, a head nod could be an agreement after a question, but during a response of the agent it is probably a backchannel signal. After the interpretation, the dialogue management component uses these interpretations to update the current state of the conversation and its participants. With the updated state, it can then decide what behaviour to perform.

If the dialogue management component decides to perform some behaviour, it sends this to the output components. The avatar shows the head movements and facial expressions of the virtual agent. The behaviour is also sent to the audio synthesis component, which generates the speech of the agent with the text that it has to say.

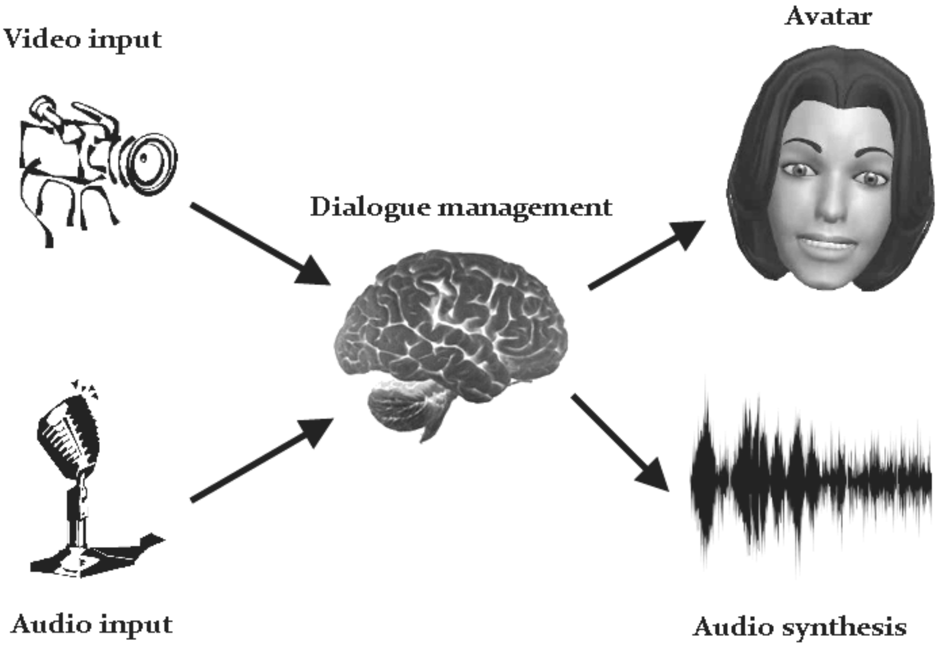

Figure 2.1: An overview of the main components of Sal.

Data is sent between components using the middleware ActiveMQ[2], and a custom-made API that enables all components to easily send and receive data to and from other components. A more detailed image of the architecture can be found in Figure 2.2. In the architecture, features are extracted, analysed and fused from the audio and video input and sent to interpreters. These modules put the interpretations of the user's behaviour and their effects on the agent into corresponding states. The action proposers use these states to suggest agent actions to perform, and send these to the output component. This component selects the actual action to perform (in case there are conflicting actions), generates the corresponding behaviour and sends this to the player which performs it with the avatar.

2.2.2 Component: Audio Input

For the audio input, the tool openSMILE[3] (Eyben et al., 2010), the core component of openEAR (Eyben et al., 2009), is used. SMILE stands for Speech & Music Interpretation by Large-space Extraction, and unites features from speech and music domains. OpenSMILE is open-source, real-time and provides incremental processing, which means that features that take longer to process — for example speech recognition — are continuously calculated and updated. In the speech recognition example, this means that openSMILE sends its current best guess of what is said continuously while the user is still speaking.

At the lowest level, openSMILE can detect a lot of musical and speech features

[2] http://activemq.apache.org/

[3] opensmile.sourceforge.net

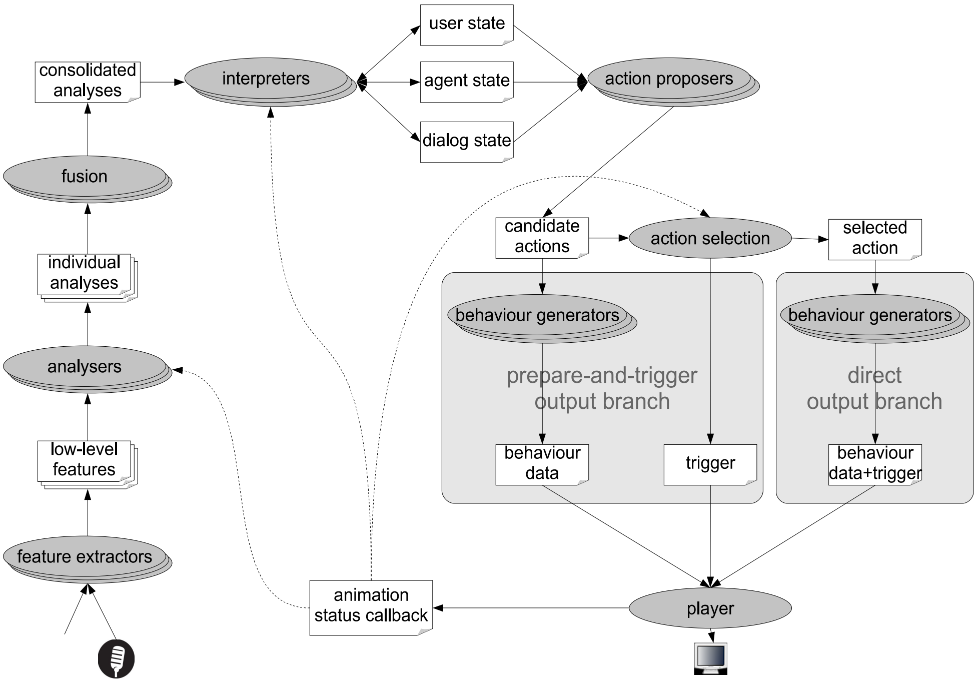

Figure 2.2: A more detailed overview of the architecture of Sal. (Schröder et al., 2010)

such as the energy of the signal, the pitch, FFT spectrum features, voice quality, and CHROMA-based features. In Sal, the low-level features that are sent to other (non audio-input) components are the fundamental frequency (F0), the probability that the current frame is voiced, and the energy of the current frame. These features are extracted and sent after each frame, which is every 10ms.

Using the features that are extracted from each frame (which are more than just F0, the voice probability and the energy), higher-level features — such as arousal, interest and whether the user thinks he has the turn or not — are calculated. These features are event-based, which means they are sent whenever they are detected with a high enough confidence. In the affective domain, the features valence, arousal, potency, and interest are calculated (Schuller et al., 2009). In the non-verbal (actually non-linguistic) domain, the features user-speaking, pitch direction, gender, and non-verbal vocalizations are calculated. The user-speaking feature sends a message whenever openSMILE thinks the user starts or stops speaking. The pitch direction is sent when there is a certain slope in the pitch, and can have the values rise, fall, rise-fall, fall-rise, high, mid, or low. The non-verbal vocalizations are either laughs or sighs of the user. Finally, openSMILE performs large vocabulary (4800 words) Automatic Speech Recognition (ASR) in an incremental way, sending updates of the recognized words each time this list changes. In earlier versions, instead of a complete ASR component a keyword spotter was used (Wollmer et al., 2009), which would detect about 150 words. The speech recognition is, amongst other things, trained on the SEMAINE Solid-Sal corpus.

2.2.3 Component: Video Input

The VideoFeatureExtractor processes the data from the video camera that records the user from the shoulders up. It sends six types of messages to other components: face presence, face location, head movement, head pose, detected Action Units, and affective features. The face presence component is straightforward: if it has detected a face in the last second it sends a message that a face is present in front of the camera. If a face is detected, the face location provides coordinates of the top-left corner of the bounding box around the face and the width and height of that box. Based on this information, it also calculates head movements: the direction of the movement and its magnitude. This data is used to detect the head nod and head shake gestures (Pantic et al., 2009). It is then possible to extract affective features from the properties of these gestures (such as direction and magnitude), as explained by Gunes and Pantic (2010b) and Gunes and Pantic (2010a).

Finally, the VideoFeatureExtractor extracts Action Units (Jiang et al., 2011). Action Units (AUs), from the Facial Action Coding System (Ekman and Friesen, 1978), specify individual or groups of facial muscles. For example, AU2 is the outer brow raiser, and AU10 is the upper lip raiser. With these Action Units, one can describe any facial expression. For example, the basic emotion happiness can be described with AU6 (cheek raiser) and AU12 (lip corner puller). This information can be used to extract useful features from the face, such as whether the user is smiling, whether the mouth is opened or not, and whether the eyebrows are raised or lowered. Also, they can be used to detect affective features. The VideoFeatureExtractor also uses the SEMAINE Solid-Sal corpus as training material.

2.2.4 Component: Avatar

The virtual agent Greta (Hartmann et al., 2002) was used to display and control the avatar. Greta consists of five major modules that form a pipeline, namely the Listener intent planner, the Listener action selector, the Behaviour planner, the Behaviour realizer, and the Player. The Listener intent planner analyses the user's behaviour when the agent is listening, and based on what the user is doing decides *when* the agent should show a backchannel signal. It accompanies this request with the communicative intention of the backchannel signal, such as whether it should show agreement or not. The Listener action selector (De Sevin and Pelachaud, 2009) receives action candidates from the Listener intent planner and from the dialogue management components. Based on the current state of the agent and the conversation, it decides which action the agent should perform. If the agent is already performing some kind of behaviour, it puts the selected behaviour in a queue, ready to start as soon as the agent has finished.

When a certain behaviour is selected to be performed by the avatar, it is sent to the Behaviour planner in a variant of FML (Function Markup Language). This component takes the characteristics of the current character and uses these to convert the received high-level behaviour to more concrete behaviour, described in BML (Behaviour Markup Language). For example, it receives the behaviour to show agreement, and using a lexicon with all communicative intentions and possible behaviours it converts this to a head nod. This BML message is sent to the Behaviour realizer, which converts BML to animation elements, whilst also incorporating lip synchronization using the speech timings of all syllables it receives from the audio synthesis. These animation elements are sent to the player, which plays them with the avatar.

Four different avatars were developed, to make the difference between the four characters very clear. They are shown in Figure 2.3.

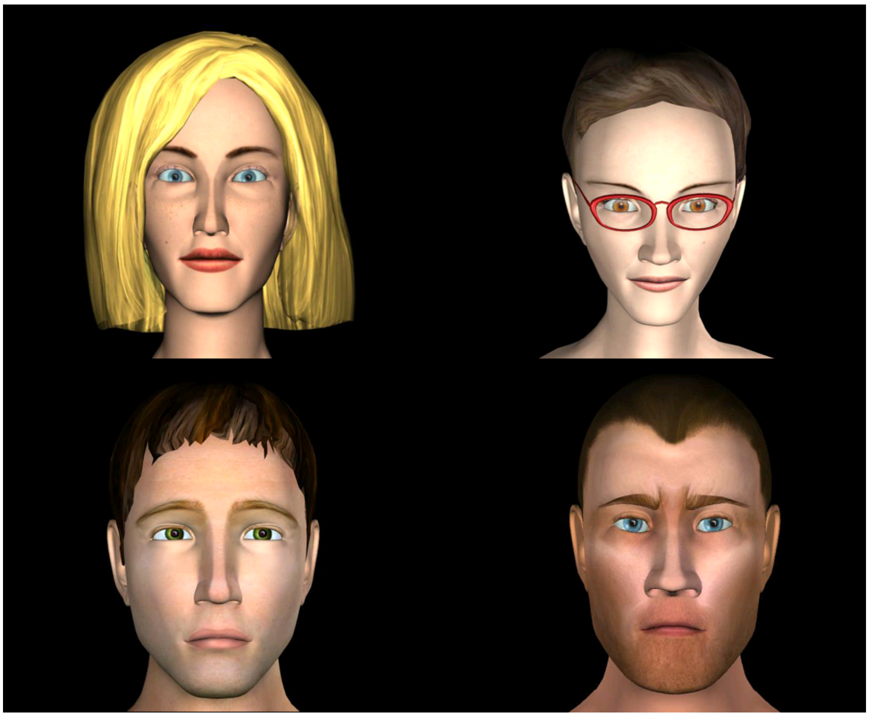

Figure 2.3: The four Sal characters as used in SEMAINE. Clockwise, starting in the top left are Poppy, Prudence, Spike, and Obadiah.

2.2.5 Component: Audio synthesis

For the audio synthesis, the open-source text-to-speech platform MARY ((Modular Architecture for Research on speech sYnthesis) was used (Schröder and Trouvain, 2003). MARY is responsible for converting the text of the agent's next utterance to sound, including speech timings for the avatar to get the lip synchronization correct. If necessary, the created voice can be manipulated by specifying accents and pitch curves.

The four characters also have different voices that match their character. For this, four professional actors were hired, based on how well their voice suited a certain character. The actors then read about 150 sentences from the respective character's script, and 500 to 2000 random sentences from Wikipedia to optimize phonetic and prosodic coverage. When generating speech, MARY works with unit-selection: it tries to find the largest matching fragment of the target-text in its database. For example, if it can find a speech element for the complete sentence in its database it will use this, but if it cannot it will try smaller and smaller blocks such as several words, single words, or even syllables. However, the smaller the fragments, the lower the quality of the generated speech. This results in voices that work well when the sentences are inside the Sal-domain, but produces sentences with a reduced quality (although still reasonable) if the text is out-of-domain.

Additionally, 30 minutes of free dialogue was also recorded with each actor, in order to extract listener vocalization (Pammi and Schröder, 2009). During these free dialogues, the actors were encouraged to be the listener as often as possible, and to use mostly "small sounds that are not words". The recorded data was used to get character-specific backchannel voices.

2.2.6 Component: SEMAINE API

All components in the SEMAINE system are coupled by means of the middleware ActiveMQ. This is a messaging server that enables components to connect to it and join Topics. If a message is sent to the server on a certain Topic, then the server broadcasts this message to all clients that have joined that same topic. This has multiple advantages: the sender of the message does not care whether a component is listening or not, it just sends its message, which makes the components more modular. Also, it is very easy to add a new component to the system: it only has to listen to the Topics it wants data from. Another big advantage is that it is easy to distribute all components across multiple computers, as long as they can connect to the ActiveMQ server. This was actually the main reason to use such a system, because at that time all components could not run smoothly together on one computer.

The next question is: what kind of messages should be sent? The messages should be easily readable by the computers, non-ambiguous, but also easily extendible and contain all information that is needed. To meet these requirements, Sal uses several standard data-formats for different parts of the system. For example, it uses EMMA (Johnston, 2009) to describe the interpretations of the user's input, EmotionML (Baggia et al., 2009) to describe emotional information, BML (Kopp et al., 2006) to describe behaviour elements, and SSML (Burnett et al., 2004) to describe details for the audio synthesis.

Unfortunately, creating an XML-message from scratch is quite a hassle, and with a system such as Sal other functionalities are also needed. To accommodate for this, the SEMAINE API was created: a meta-component that is responsible for all other components and their communication. It monitors the state of all components and notices whether one is stalled or has crashed. Using a GUI, it can show all components, their current state and via which Topics they are connected. The API also takes care of easier communication using namespace-aware XML and provides methods to send and receive certain information without using XML directly — the API automatically converts the data to XML and vice-versa. The API also takes care of a centralized time, to provide each component with the same time stamp, and centralized logging.

2.3 Dialogue Management

The dialogue management component (from now on called the DM) is a complex component, responsible for interpreting the detected behaviour, keeping track of the agent's states and generating new behaviour for the agent. In this section, we elaborate on these aspects, starting with the agent's states. We continue with the interpretations that are made, the turn-taking module that decides when the agent should start speaking, and the action proposer that generates new behaviour for the agent.

2.3.1 The agent's states

The DM stores its knowledge about the user, the conversation and about itself in *states*, to be precise the UserState, the AgentState and the DialogueState. The UserState contains the 'current best guess' of the user's behaviour: the interpretation that, given the current evidence, is most likely true at this point in time. The state is filled by different interpreters, which analyse the detected behaviour of the user and decide whether there is enough evidence to make a certain interpretation. For example, if the detected gender 'Female' has a confidence value greater than 0.8, then an interpreter could add to the UserState the current best guess that the user is female.

The DialogueState contains information about the conversation itself. In systems that focus more on the content of the conversation, this could include the phase of the conversation, the current topic, etcetera. However, in Sal it only contains several variables, of which two keep track of the current turn-state. There are systems that only use one variable with two possible states to keep track of the turn, which means that either the user or the agent has the turn. However, turn is not an objective item one can possess, there is no regulating object that says that now the user or the agent has the turn. Instead, the user can think he or she has the turn or not, and the agent can think it has the turn or not. This makes it possible that both participants think they have the turn — resulting in a clash — or that both think they do not have the turn — resulting in silence. For this reason, both the user's turn-state and the agent's turn-state are stored.

The AgentState stores information about the agent, such as its emotional state, a history of performed behaviour, and its eagerness to start speaking. Modules like interpreters analyse the user's behaviour and decide how this affects the user's state. For example, a long silence might increase the agent's eagerness to start speaking, and a smiling user could increase Poppy's happiness value.

2.3.2 Interpreters

The interpretation modules are responsible for analysing the user's behaviour, decide how this could be interpreted and how it affects the different states. Sal contains the following interpreters:

- **Emotion interpreter** Responsible for putting detected user-emotions into the UserState. First, a Fusion module combines the detected emotions from the audio and the video, and if the confidence of the consolidated emotion exceeds a certain threshold, then it is put into the UserState.

- **Non-verbal interpreter** Responsible for putting detected non-verbal behaviour — such as head movements, laughter, etcetera — into the UserState, using a similar approach as the emotion interpreter. First, a Fusion module combines detected behaviour from the audio and video input components into a single event, and if the confidence of this event exceeds a certain threshold it puts it into the UserState.

- **Agent mental state interpreter** The mental state of the agent is a set of twelve communicative intentions (see Bevacqua et al. (2008)) — such as agreement, belief, interest and understanding — which are mainly used to select the type of backchannel signal to be used. In Sal, each character has a baseline: a default value for each intention. For example, by default Poppy is more agreeing and interested than Spike. When an affective state (for example valence or arousal) is detected, this affects the agent's mental state, based on the emotional state of the current character. If the detected affective state is congruent with the agent's emotional state, then intentions such as agreement and liking will increase, and vice versa. For example, if Poppy detects a high valence or arousal with the user (which matches its own emotional state), then its mental state changes in favour of the user, and after that Poppy will be more agreeing and liking.

- **Turn-taking interpreter** The turn-taking interpreter is responsible for deciding when the agent should take the turn, based on the user's behaviour. We will discuss this module more extensively in the next section.

2.3.3 Turn-taking

The turn-taking module of Sal is responsible for deciding when the Agent should start speaking, based on the user's behaviour. In earlier versions of Sal, only the speaking behaviour of the user was considered; whenever the user was silent for more than two seconds, the agent would take the turn. This value was chosen to make sure that the agent's response would not overlap with the user's turn. However, this is very reactive behaviour, and for fluent conversations this is just not acceptable.

In Sal, the turn-taking is not only reactive, but also proactive, by keeping track of its own intention to speak (how eager it is to take the turn). This intention changes according to the user's behaviour, but also to other factors such as time. If the intention to speak gets high enough, then the agent starts speaking. The observation that the user has finished a turn contributes to this intention, but this is not compulsory; if other factors increase the intention to speak enough, then the agent will also start speaking if the user has not yet finished.

To calculate the intention to speak, the following aspects are used:

- **User silence time** A value between 0 and 100 which increases over time when the user is silent. When the user starts speaking again this value drops back to 0.

- **User's emotion** A value between 0 and 80 with 10 points for each detected emotional event (for example a peak in the arousal). This encourages the agent to respond faster if the user shows more emotion, because this also means the agent has more to respond to.

- **User speaking time** A value between 0 and 30 that increases over time when the user is speaking (reaches its max after 30 seconds). This is to stimulate the agent to take the turn if the user has already been speaking for a longer period.
- **Agent turn-end wait time** A value between -100 and 0 that starts at -100 after the agent finishes its turn, and for the next two seconds rises to 0. This makes sure that the agent does not start too soon after its own utterance.
- **User not responding** A value between 0 and 100 that starts rising if the user does not start talking after the agent finishes its turn. It starts after 2 seconds and rises to 100 in 4 seconds unless the user starts speaking. This makes sure that the agent takes the turn if the user does not start his or her turn.

The intention to speak is calculated by adding these values together, resulting in a value somewhere between 0 and 100. For each character, a certain threshold determines when that character takes the turn. Based on the results of this thesis (see Chapter 3 to 6) we gave the characters different thresholds. For example, Poppy's and Spike's thresholds are fairly low, and Obadiah's threshold is high, which means that Poppy and Spike react much faster than Obadiah. This is in line with their emotional characteristics, since Poppy and Spike are meant to be more aroused than Obadiah.

2.3.4 Action proposer

The action proposer is responsible for selecting what to say when the turn-taking module decides the agent should say something. Chapter 7 and 8 explain in more detail how a response is selected, but this section roughly explains the process.

The action proposer uses a set of templates to select a response, based on the current phase of the conversation and the user's behaviour. It uses features such as detected affective states (e.g. valence and arousal), non-verbal behaviour (e.g. head nods, laughs), and low-level audio features (e.g. pitch and energy). It continuously processes these features to determine what is the best response at that moment. If the agent wants the turn, then this response is selected to be performed, but if the agent is still listening, then the best response is prepared in case the agent takes the turn soon.

When determining what to say, the action proposer uses several strategies. At the start of the conversation it uses a predetermined script of three responses, in which the agent introduces itself and asks how the user is today. If the user or the agent wants to change the character, then it follows a script too, in which it tries to determine with the user which character it should change to. In between these phases, it uses templates to determine which response or what type of response it should give. Such a response type could be, for example, the introduction of a new topic, to cheer the user up, to insult the user, or to ask for more information. If no template matches, and no good response can be found, it falls back to a generic response that fits most situations, such as 'Tell me more'.

The templates are implemented using Flipper[4], a specification language and interpreter for Information-State update rules (Ter Maat and Heylen, 2011). Such an

[4] http://sourceforge.net/projects/hmiflipper/

update rule describes the behaviour to be performed when the rule is fired, the effects it has on the state of the conversation, and the current conversation state that is required to fire the rule. An example of such a template is this:

```
<template id="RespondToSmile1" name="A response to a smile">
 <preconditions>
  <compare value1="$face.nrOfSmiles" comparator="greater_than" value2="1" />
  <compare value1="$speakingIntention" value2="want_turn" />
 </preconditions>
 <effects>
  <update name="$nrResponses" value="$Agent.totalResponses + 1" />
  <update name="$responses._addlast" value="#Response129" />
 </effects>
 <behaviour class="ActionProposer" quality="0.5" />
  <argument name="response_id" value="#Response129" />
 </behaviour>
</template>
```

This template checks if there was a smile and if the agent wants the turn. If so, then the total number of the agent's responses is incremented by one, the agent's response (#129) is put in a list of performed responses and the selected behaviour is executed.

2.4 Sal Evalution

An important aspect of designing a system such as Sal is the evaluation afterwards. Most components were tested during their development, but evaluating the complete system is a challenge. This section explains roughly how Sal was evaluated and with what results, and more details can be found in Schröder et al. (2011).

2.4.1 Evaluation methods

Evaluating Sal as a complete system was a challenge, mainly because it is not a usual system. For example, there are no tasks to be achieved, which means that measurements such as effectiveness and efficiency could not really be used. Also, negative affective states such as frustration are sometimes good; if users get frustrated by Obadiah's eternal depression it actually means that they are engaged in the conversation, which is precisely what we wanted: to measure the user's engagement.

In order to evaluate Sal, three methods were used:

1. **Questionnaire** A simple way to measure engagement is to ask the user. However, asking things during the conversation disrupts the flow the user might have with the agent, and asking afterwards causes the user to rationalize things. Therefore, we implemented a small questionnaire in the system itself, as a neutral fifth character. This character appears each time the character is changed, and asks the following three questions of the user:

    (a) How often did you feel the avatar said things completely out of place? (appropriateness)

    (b) How much did you feel that you were involved in the conversation? (felt engagement)

(c) How naturally do you feel the conversation flowed? (flow)

These questions focus on the three most important elements: the agent, the user, and the conversation.

2. **Yuk button** In order to get more detailed feedback during the conversations, we asked the users to hold a button and press it every time they felt that the simulation was not working well. This provided a non-verbal measure of how well the system was working, but also details about problem areas.

3. **Annotated engagement** As a final measure, an annotator watched the conversations as they took place, and annotated the user's apparent engagement using a trace-like technique.

2.4.2 Evaluation results

When evaluating a system, you need a baseline to compare it with. However, there is no system like Sal, which means there is no baseline either. Therefore, the final system was compared with a version of the system in which affective features of the output were disabled; the voices and faces were expressionless and the agents did not produce backchannel signals. In total, 30 users took part in the evaluation, of which 24 female and 6 male. The users talked with both versions of the system (in counter-balanced order), and with each version they talked with all characters in a random order.

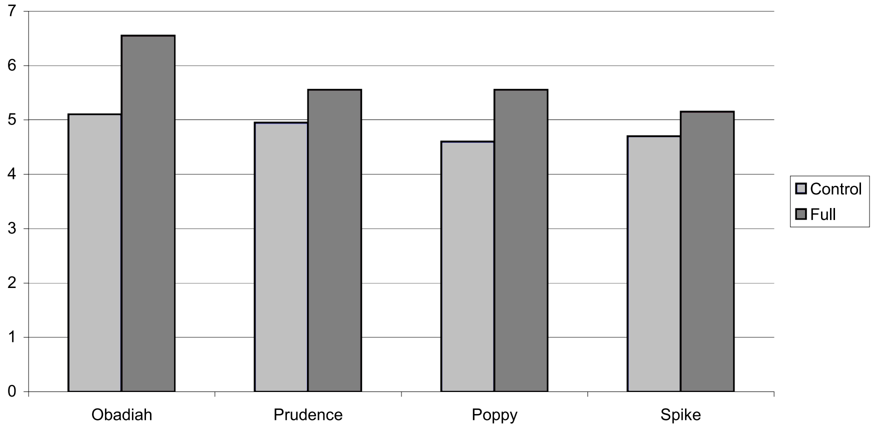

Figure 2.4: Mean values for self-reported Flow and Felt engagement.

Figure 2.4 shows the mean values for self-reported flow and felt engagement. This graph shows that the full system scores significantly ($F(1,29) = 9.25, p = 0.005$) higher than the control system. Figure 2.5 shows that in the full system the users were perceived as significantly ($F(1,29) = 23.3, p < 0.001$) more engaged than in the control system.

To give a feeling of how well the automatic system worked compared with the semi-automatic version — in which the users thought they were talking with the real system, but a wizard controlled the response selection — Figure 2.6 shows the mean

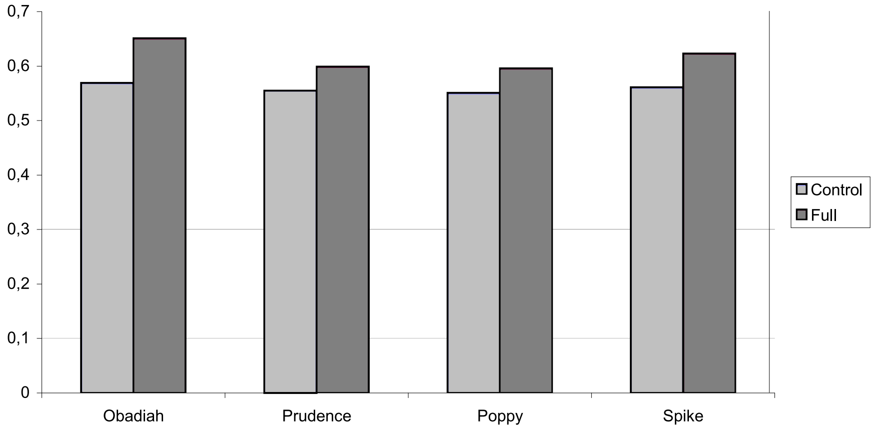

Figure 2.5: Mean values for annotated engagement.

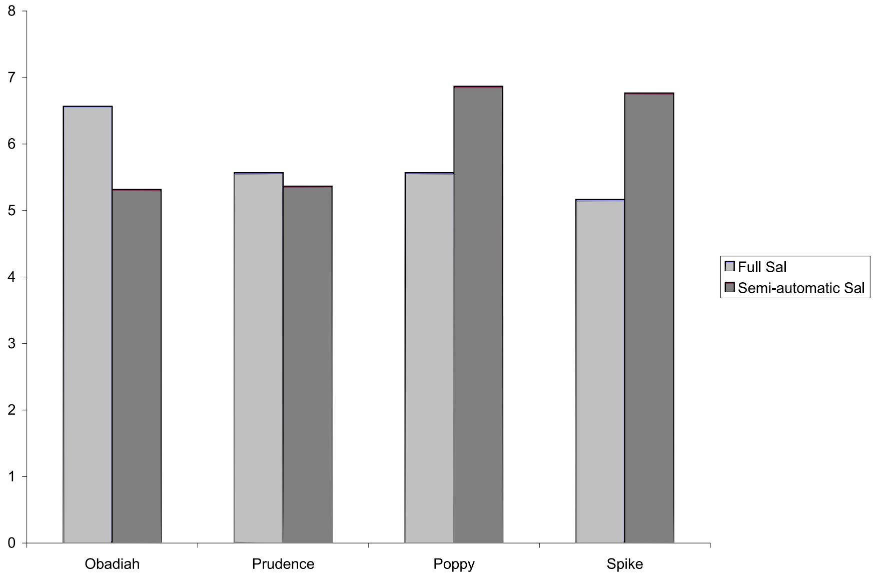

Figure 2.6: A comparison of the Full version and the semi-automatic version of Sal for Flow and Felt engagement.

flow and felt engagement of both tests. It shows that the automatic system works well for Obadiah and Prudence, but not at all for Poppy and Spike. It also shows that some characters are evaluated better in the full system, and this is a key point in the evaluation: the character has a big effect on the results.

2.5 Conclusion

In this section, we explained the context of this thesis, namely the SEMAINE project. We explained several reasons why we need sensitive artificial listeners, for example to have a system that can induce emotions in a natural conversation setting. We explained that the system we created in SEMAINE, a virtual listening agent system called Sal, consists of video and audio input, dialogue management, and an avatar with speech. We elaborated on these components, especially on the dialogue management, and finally we briefly explained how Sal was evaluated.

The remainder of this thesis is divided into two parts. The first part is about how we can use turn-taking behaviour as a tool to amplify the differences between the different characters. We will present two studies in which we tried to find out how people perceive different turn-taking strategies, and we think about how we can use these results to create different impressions. In the second part of this thesis, we will describe our work on response selection. We will explain which methods we used to select an appropriate response from the fixed list of responses, using only the detected non-verbal behaviour.

Part II

Turn-Taking

3 Turn-taking and perception

This part of this thesis is about how people perceive different turn-taking strategies. The aim is to use this knowledge in virtual agents, to use turn-taking as a tool to change the impression that people have of the agent. In this chapter, we will provide an overview of turn-taking in the literature, social signals that can be found in turn-taking behaviour, and why we should use turn-taking as a tool. In Chapter 4, we will present our first study, in which we had people rate unintelligible conversations in which one participant uses different turn-taking strategies. In Chapter 5, we will present our second study, in which the rater actively participated in a conversation with a wizard-controlled agent that used different turn-taking behaviour. Finally, in Chapter 6 we will provide the conclusions and a reflection about these two studies.

3.1 Turn-taking in the literature

How can something that seems so simple and natural to humans be so complex? For humans, turn-taking is something we do without thinking. To quote Yngve (1970):

> When two people are engaged in conversation, they generally take turns. First one person holds the floor, then the other. The passing of the turn from one party to another is nearly the most obvious aspect of conversation.

However, even after more than 40 years of research on this phenomenon, we are still unable to build a virtual character that has the same smooth turn-taking capabilities as humans.

Yngve (1970) notes that although the turn passes from one party to another, having the turn or not is not the same as being the speaker or the listener. He argues that it is possible to speak out of turn, which even happens reasonably frequently. According to Yngve, "...both the person who has the turn and his partner are simultaneously engaged in both speaking and listening". He does not mean that both interlocutors simultaneously have the turn, but that the person who does not have the turn can send messages on what Yngve calls the back channel. On this channel, the person who has the turn receives short feedback messages. These messages can be small comments or nods, but also longer comments such as “Oh, I can believe it” or even questions such as “You’ve started writing it then — your dissertation?”.

Sacks et al. (1974) followed with their famous paper, in which they explain three simple turn-taking rules that are followed after each turn, while trying to minimize the duration of silence between turns and the duration of any overlapping speech. At the end of a turn, the following three rules are used:

1. If the current speaker selected the next speaker, then the selected speaker has the right and is obliged to take the turn, and the other participants do not.

2. If the current speaker did not select the next speaker, then a participant may select him or herself as the next speaker, and the person who starts first acquires the turn.

3. If no participant self-selects him or herself as the next speaker, then the current speaker may continue to speak.

Simple rules, but unfortunately, there are some problems with them. A lot of these problems are explained by O’Connell et al. (1990), who criticize the assumptions, concepts, and methods of the paper by Sacks et al. (1974). For example, they have a lot of criticism on the fact that Sacks et al. only provide anecdotal evidence. This makes the conclusions weak, and easy to oppose by offering counter examples. This is amplified by the fact that, in general, time (for example durations of silence or overlapping speech) is often treated intuitively and perceptually instead of objectively by measuring it. Also, they contest the assumption “Someone’s turn must always and exclusively be in progress.” Instead, they argue, at any time in a conversation the turn belongs to *all* participants. Turns may remain unclaimed, and the fact that a user interrupts someone else does not by itself make the conversation faulty.

Something to keep in mind with the rules of Sacks et al. is that it is a *descriptive* model of turn-taking: it tries to create a model for turn-taking that explains the turn-taking behaviour that is perceived in recorded conversations. Another example of a descriptive model is provided by Padilha and Carletta (2003). In this paper the authors describe a multi-agent system that simulates a discussion, and they want these discussions to statistically approximate human-human discussions. Their turn-taking model uses additional non-verbal behaviour such as gaze, gestures, posture shifts, and head and facial expressions to decide what to do.

However, for a virtual agent that uses a turn-taking model to determine how to behave, we do not need a descriptive model, but a *generative* model. A model that does not try to explain perceived behaviour, but can predict and generate turn-taking behaviour. A model that analyses the user’s behaviour, and can tell whether a small silence is just a pause or the end of a turn, and whether a user’s ‘hmm’ is just a backchannel signal or an attempt to take the turn.

Most studies on generative models focus on predicting (or detecting) when a user finishes his or her turn. For example, Sato et al. (2002) describe how they created a decision tree that, given a silence longer than 750 ms, determines whether it is only a pause of the user or a possibility for an agent to take the turn. The classifier was tested by comparing its behaviour to a baseline system which would take the turn after each silence that was longer than 750 ms. To be able to respond faster, Raux and Eskenazi (2008) developed a system that dynamically changes the time it waits after a user silence before it takes the turn, based on automatically extracted features of the audio. These algorithms focus on determining when the end of a turn is reached as fast as possible .

However, being able to predict or detect when a user has finished his or her turn is not enough, something needs to be done with this information. A great example of this is the Ymir Turn-taking Model (YTTM) of Thórisson (2002): a turn-taking model that addresses all aspects of turn-taking, namely multi-modal perception, knowledge representation, decision making and action generation. In this model, multi-modal input detected from the user flows into three different layers, which all operate at a different frequency: the reactive layer, the process control layer, and the content layer. For example, the fastest layer — the REACTIVE LAYER — has a perception-action loop of two to ten times per second, and is responsible for highly reactive behaviour, such as gazing at mentioned objects. During each perception-action loop (which runs at a different speed for each layer), each layer passes its input through a set of perceptors to interpret the input. These interpretations are then fed through a finite state machine that keeps track of the current Turn status (who has the turn and who wants, takes or gives it). Changes in this state machine can affect decision rules, which can fire certain behaviour of the agent. For an overview of the layers of the YTTM, see Figure 3.1.

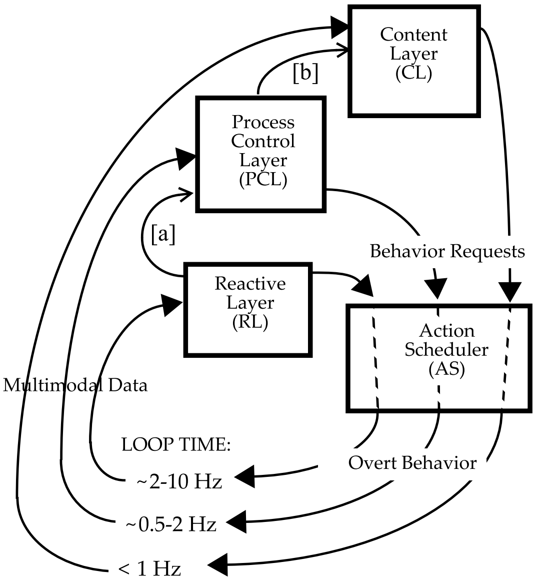

Figure 3.1: The three layers of the YTTM, image taken from Thórisson (2002). The loop times are the frequencies of the full perception-action loop. [a] and [b] are partially processed data.

The YTTM system uses a single notion of TURN: either the user or the agent has the turn. This implies that the TURN is a property of the conversation. But Yngve (1970) already noted that “each person in a conversation acts according to his own concept of who has the turn". With a single notion of turn, situations where both or neither participants think they have the turn cannot be modelled.

3.2 Optimal Turn-taking

Something that Thórisson (2002), Raux and Eskenazi (2008) and Sato et al. (2002) all tried to achieve, whether implicitly or explicitly, is *optimal* turn-taking behaviour. They all wanted turn-taking to be as smooth as possible, with as little overlap as possible and the silences between turns as short as possible. Another clear example of this is the work of Jonsdottir et al. (2008), in which the authors used machine learning techniques to have an agent learn 'proper' turn-taking behaviour. They did this by creating a classifier that uses the prosody of the other person's speech to determine how long to wait before taking the turn. They defined 'proper' behaviour as turn-switches with no overlap and as short a gap between the turns as possible, and they evaluated the system on these properties. When prosodic information is not enough to determine whether a pause is the end of a turn, other non-verbal signals might help (as shown by Barkhuysen et al. (2008)). For example, Novick et al. (1996) show how gaze behaviour is used by humans when managing turn-taking.

The reason that people are so focussed on smooth turn-taking can probably be traced back to the paper by Sacks et al. (1974). In this paper, the authors specifically provide rules for turn-taking in order to "minimize gap and overlap", where a gap is defined as the silence period between two consecutive turns. This view is strengthened by studies such as De Ruiter et al. (2006), who show that 45% of the speaker transitions in their recordings of telephone conversations had a floor transfer offset — the start time of the next turn minus the end time of the previous turn — of between -250 and 250 milliseconds, and 85% between -750 and 750 milliseconds.

But why focus on smooth turn-taking? O'Connell et al. (1990) argues that the turn-taking rules of Sacks et al. (1974) are entirely focussed on the 'smooth transitions', but they do not provide sufficient evidence that 'the vast majority of transitions' is smooth. Also, according to Orestrom (1983), evidence indicates that "speaker-shift is seldom, if ever, an entirely smooth process". And O'Connell et al. continue with "There is no evidence that there is only one best way or any ideal way to carry on [a] conversation, either in terms of mandatory turn allocation or in terms of mandatory time allocation". As an example, look at the graphical representation of a telephone conversation in Figure 3.2 (image taken from Campbell (2009)). This figure shows the speech on/off behaviour of two people talking to each other using a telephone. Observe that the figure shows high degrees of overlapping speech, and the concepts of 'turn' and 'utterance' are hard to find.

There is nothing intrinsically wrong with overlapping speech. Schegloff (2000) provides an insightful overview of overlapping speech in conversations. In his paper, he gives a lot of examples of conversations in which overlapping speech occurs, and he summarizes the methods that are used to resolve the situation. Many overlaps are resolved after a single syllable, when one of the two participants withdraws. Longer overlaps are usually stopped within one beat after one of the speakers switches to COMPETITIVE SPEECH, that is, talking louder to keep the turn.

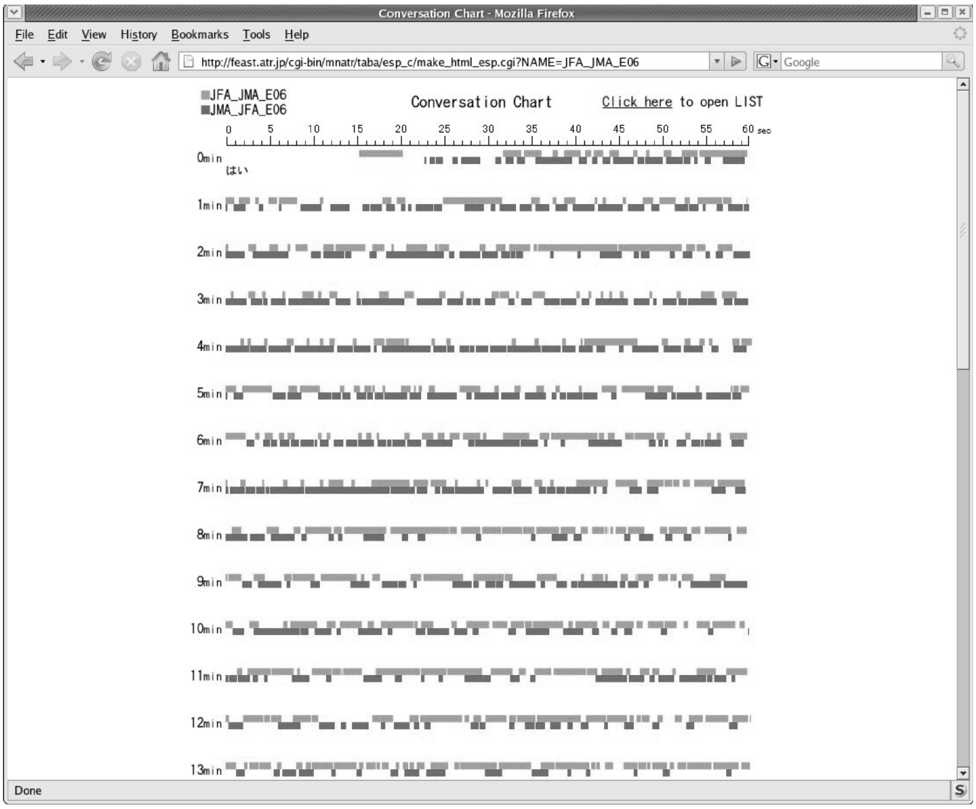
Conversation Chart - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://feast.atr.jp/cgi-bin/mnatr/taba/esp_c/make_html_esp.cgi?NAME=JFA_JMA_E06
Google
JFA_JMA_E06
JMA_JFA_E06
Conversation Chart
Click here to open LIST
0 5 10 15 20 25 30 35 40 45 50 55 60 sec
0min
はい
1min
2min
3min
4min
5min
6min
7min
8min
9min
10min
11min
12min
13min
Done

Figure 3.2: An example conversation, taken from Campbell (2009)

The overlapping speech gets *resolved*, which indicates that in those cases it was unwanted. But interestingly, in his analysis Schegloff excludes four types of overlaps which occur regularly in conversations, "primarily those types in which the simultaneous speakers do not appear to be contesting or even alternative claimants for a turn space. In these cases, the conduct of the participants does not show these occurrences to be taken as problematic by them". Basically, these are types of overlaps that are non-problematic and acceptable. These are the four cases:

1. **Terminal overlaps** – The next speaker starts just before the prior speaker finishes, usually because the end of that turn is noticeably getting near. Note that this is exactly the type of overlapping speech that a lot of researchers try to avoid at all costs, but Schegloff argues that this case of overlap is not problematic.

2. **Continuers** – These are the same as what Yngve (1970) calls backchannel signals: short signals to indicate that you are still following the speaker, such as 'uh huh', 'hmm', and 'yeah'.

3. **Invited overlapping speech** – Cases in which the speaker invites another participant to speak, without yielding the turn. For example, inviting the other participant to find a word that the current speaker cannot retrieve at that moment, or collaborative utterance construction in which the other participant is invited to finish it.

4. **Choral talk** – Forms of talk that all participants are required to do simultaneously, for example laughing, greeting, and congratulating.

In short, Schegloff shows that if a (potentially undesirable) overlap occurs, this is usually resolved very swiftly, without hampering the conversation too much. Also, he argues that a lot of cases of overlapping speech are not problematic at all.

Ten Bosch et al. (2004) also show that overlapping speech is indeed very common. They define different types of turn changes, and extracted these types from two corpora: one containing face-to-face conversations and one containing telephone conversations. In total, almost 20,000 turn changes where analysed. In face-to-face conversations, 80% of the turn transitions appear after a small pause after the end of the previous turn, but in 20% of the cases the next turn starts while the previous turn was not finished yet. In telephone conversations, the percentage of early turn starts (which overlap the previous turn) increases to 26%.

Ten Bosch and Oostdijk (2004) show that longer pauses are also not uncommon. In their paper, the authors analysed 93 telephone conversations of two people who knew each other well and could talk about anything they wanted. The conversations lasted about nine minutes on average. In 15 of these conversations at least one of the interlocutors had an *average* pause-duration (between and within turns) of at least 0.5 seconds. In seven of these 15 conversations, both participants had an average pause-duration of at least 0.5 seconds. This shows that, short (less than 0.5 seconds) pauses are most common, but longer pauses (within and between turns) happen regularly too.

These sources and examples show that ‘smooth’ turn-taking is not always as common as some people believe. Turn changes with overlapping speech occur regularly, and often without causing any problems to the conversation. This does not mean that one should not focus on turn-taking without overlapping speech and with short gaps, but one should realize that this is not the only turn-taking behaviour that occurs in the real world, and that there are cases in which an overlap or a long gap is acceptable and that should also be modeled to create realistic virtual humans.

3.3 Social signals in turn-taking

Now that we have established that turn-taking does not have to be smooth *per se*, it is time to go one step further: what does a speaker’s turn-taking behaviour tell us about his or her social or affective attitude or state? Turn-taking behaviour can be very informative, and different situations — for example, type of conversation, phase of the conversation, status of the participants, and emotional state of the participants — can result in different turn-taking behaviour. For example, Trimboli and Walker (1984) look at differences in turn-taking behaviour in cooperative and competitive conversations. In the cooperative scenario, the participants held a friendly conversation, and in the competitive scenario, the participants had to hold an argument on a topic they chose themselves. They found that the cooperative conversations had significantly less interruptions than competitive conversations. Orestrom (1983) also noticed that “A clash of opinions also means a clash of turn-taking”.

Goldberg (1990) extensively describes interruptions in conversations. She explains that interruptions are often linked to displays of power and control, and are generally viewed as rude and disrespectful. However, she also lists studies that show that this correlation between interruptions and displays of power and control is weaker than previously presumed. She argues that interruptions may also convey rapport or cooperation, and can be triggered by the participant's enthusiastic interest and active involvement. Some studies found that interruption frequency varies with respect to the current phase of the conversation, and the 'meaning' of that interruption changes with it. This means that assigning personal or relational attributes to a participant based only on the number of interruptions is unsound; the context of the conversation needs to be taken into account too. Although, Rienks and Heylen (2006) show that the number of interruptions can help when determining dominance in a conversation. They show that the number of successful interruptions is one of the top discriminative features for dominance in meetings.

Goldberg studied interruptions, and tried to differentiate between power and non-power (rapport) interruptions. Power interruptions are generally perceived as more rude, impolite, intrusive and inappropriate, while rapport interruptions are generally understood as expressions of open empathy, affection, solidarity, interest, etcetera. She found that power-interruptions are usually off-topic, introduce new topics, or contain little relevance to the previous subject. Rapport interruptions have a greater degree of 'fit' (as she calls it) with the interrupted utterance.

Goldberg's literature review is mostly about what interruptions tell about someone's personality. Instead, Robinson and Reis (1989) look at how interruptions change the perception of a listener. They let subjects listen to a four-minute audiotape of a discussion between two interlocutors. The subjects had to rate these interlocutors on masculinity, femininity, competence, sociability, attractiveness, and traditionality. They varied the sex of the interrupted interlocutor, the style of the interruption (statement, question, or no interruption), and the status of the conversational interlocutor. They found that interrupters were perceived as more assertive, more masculine, and less sociable. In contrast, interlocutors who were interrupted were seen as more passive and submissive. They did not find differences between interrupting with a statement or a question.

Beattie (1981) studied differences between interruption behaviour and status. He analysed videos of tutorial groups —containing discussions of a tutor and several students — for interruption behaviour of the participants. He found that high-status participants (the tutors) were interrupted significantly more often than low-status participants (the students). At first sight, this seemed the opposite of what the author expected, but he argues that this is because the students are trying to make a good impression on the tutor, which causes them to interrupt more often to state their view. The author also shows that students' interruptions are significantly more often barge-in interruptions, in which the current speaker has not reached a possible completion point yet. Mostly likely, this was caused by the competitiveness between the students. When tutors interrupt, they generally use overlapping interruptions just before the end of a turn is reached.

Trimboli and Walker (1984) compared turn-taking styles of cooperative and competitive conversations. They found that in cooperative conversations there were significantly less interruptions. However, they also looked at another interesting feature of turn-taking: silences. They found that in cooperative conversations there were significantly longer pauses within each turn. But that is just one example of how informative silences can be. According to Clark (1996), cessations and pauses are signals, and are powerful cues for what is happening in a conversation. Pauses can be politeness markers (Brown and Levinson, 1987) and can be used to indicate cognitive processing, for control mechanisms (to control the flow of the conversation), to indicate acceptance and refusal, and for turn-taking (Endrass et al., 2008).

These examples make clear that specific turn-taking behaviour — for example starting while the other person is still talking, or leaving longer periods of silence between the turns — can be very informative, for example about personality, mental state, politeness, and other social factors, and therefore can be used in the design of virtual agents to create a distinct persona.

3.4 Virtual Agents and Personality

In the previous sections we discussed personality and emotions of people. Whether explicitly designed or not, virtual characters also have a certain 'personality' and they convey a certain emotional state. For most characters this is the 'extremely polite' personality, and in most cases this is a very appropriate attitude. For example, for agents that function as a virtual receptionist (e.g. Mack (Cassell et al., 2002) or MIKI (McCauley and D'Mello, 2006)) or a museum guide (e.g. Max (Kopp et al., 2005)), it is their responsibility to address visitors as politely as possible. However, as already stated in Chapter 2, there are sometimes good reasons to change this personality. For example, virtual actors playing a certain role, personal virtual agents that can become angry or sad based on what you say, etcetera. An important question is how to make the agent's behaviour, looks and speech match its personality and emotional state.

To have a virtual agent show its emotional state, changing its facial expression is an obvious choice (Arya et al., 2009; Mancini et al., 2007). But besides facial expressions, the human face has other options to display a certain emotion. For example, crying is a very powerful cue to the emotional state, and Van Tol and Egges (2009) explain how virtual agents can cry too. De Melo and Gratch (2009) describe even more facial signals that a virtual character can show besides moving muscles, namely wrinkling the forehead, blushing, and sweating.

Another modality that can be used to show a certain personality or emotion is speech. Without now considering *what* the agent is saying, *how* the agent says it is very important. Türk and Schröder (2008) study three different methods to transform neutral speech — generated with unit selection from a large neutral database — into emotional speech, in this case cheerful, aggressive, or depressed. These methods modify the prosody of the neutral speech that is generated first. Theune et al. (2006) define rules to modify neutral speech to a 'storytelling' speech style by changing the pitch, the intensity, the speaking tempo, durations of vowels, and pause lengths. Mozziconacci (1998) shows that pitch level, pitch range and speech rate can be used to convey different emotions too.

Changing the agent's bachchannel behaviour is also a great way to show different emotional states or personalities. For example, the different characters of Sal use different types of backchannels, and these types are changed by the current emotional state of the agent too (Bevacqua et al., 2010). The method used is explained in more detail by De Sevin et al. (2010). In their paper the authors explain how personality influences the type and the frequency of the agent's backchannels and mimicry. In the ECA developed in the Companions project (Smith et al., 2010), the agent modifies its backchannels based on the detected emotional state of the user instead of its own state. But this method can be used to change the perception of the agent too, for example by making the agent seem more empathic by performing more affective backchannels.

Finally, another modality that can be varied is gestures. Von der Pütten et al. (2010) show that differences in gesture rate and gesture performance parameters significantly change the perception of extraversion.

3.5 Turn-taking as a tool

In the previous section, we showed that there are numerous ways to give a virtual agent a certain 'personality' or appearance. In section 3.3 we showed how aspects such as personality, status, perception, and other social aspects are related to different types of turn-taking behaviour. So, why not combine these two elements, and use turn-taking as a tool to adjust these aspects? For example, say, you want a more assertive virtual character. This could be achieved by interrupting the user more often (Robinson and Reis, 1989). And enthusiasm could be augmented with faster turn-taking behaviour (Goldberg, 1990).

However, most of these examples are taken from studies which have looked at the correlation between emotion, personality and other social factors, and measured turn-taking behaviour. For example, this means that in a competitive conversation more interruptions are observed. But this does not automatically mean that more interruptions make the conversation more competitive.

What we need to know is how different turn-taking behaviour influences the conversation or, more specifically, influences the perception of the participants. How do the participants perceive the agent when it uses different types of turn-taking?

In this part of the thesis we will study this. In Chapter 4 we will describe the PASSIVE experiment, in which we automatically generated incomprehensible conversations in which one participant varied the start time of its utterances, and its behaviour when overlapping speech is detected. Using a questionnaire we studied how humans perceived this participant on scales concerning personality, emotion and interpersonal stance.

In Chapter 5 we will describe the ACTIVE study, in which we let the user actively participate in the conversation, instead of being a bystander. Using a Wizard of Oz setup, the agent acted as an interviewer and asked the user a number of questions. The wizard controlled the start time of the questions, and based on the strategy started it early, directly when the user finisheed his or her answer, or late. After each interview, the user completed a questionnaire about how he or she perceived the agent, again on different scales concerning personality, emotion and interpersonal stance.

In Chapter 6 we will summarize and discuss the results of the two studies. We will look at differences and similarities in the results, and how the results match with the literature. We will also discuss how the results can be used practically in virtual agents, and we will conclude with possible directions for future work.

4

Turn-taking perception when listening to unintelligible conversations

This chapter is about the first study we performed to find out how turn-taking affects perception (Ter Maat and Heylen, 2009). More precisely, we wanted to find out how variations in turn management strategies affect how people perceive aspects of personality, emotional state and interpersonal stance of the person (or agent) that uses said strategy. This means we needed to show, in some form, these different turn-taking strategies to human participants, and give them the opportunity to write down their perceptions.

However, a problem with measuring perception is that a lot of factors are involved. In order to compare various turn-taking strategies, in the ideal situation all other aspects remain the same. Therefore, in the first study, called the PASSIVE STUDY, we decided to let the participants listen to prerecorded conversations. To maximise our control on the conversations and the start-times of the turns, we used a computer system to generate several conversations where one agent talked to another, varying the ways in which one agent dealt with overlap and when it started speaking. These conversations were presented to human raters who judged the agent on various semantic differential scales. To generate the conversations we built a conversation simulator that allowed one to define different turn management rules. Before we specify the turn management strategies and the impression measures that we examined, we will first present the conversation simulator.

4.1 Conversation Simulator

The conversation simulator (Ter Maat and Heylen, 2010) allows one to program the behaviour of two agents that can communicate with one another by sending each other information on their communicative actions. They indicate whether they are silent, speaking normally or using elevated speech. These actions, however, are subject to a small delay of 200 ms before they are recognized by the other agent, to simulate the fact that humans need a bit of processing time as well before they interpret what they see or hear.

Furthermore, to be able to define more refined rules for regulating turns, the agents can also send certain signals that indicate what they are about to do. The agents can send a signal that indicates that the end of an utterance or a turn is near, that the agent will start a turn soon, or a signal that the agent wants to keep the turn. These signals are sent by humans as well, for example as described by Duncan and Niederehe (1974). These signals are sent directly to the other agent, without any delay and without any interpretation-steps. However, to simulate that turn-taking signals can be ambiguous and misinterpreted by the recipient, a certain error margin has been introduced. There is a 10% chance that signals are removed before reaching the other agent, and there is a 20% chance that the meaning of the signal is changed. This should add enough noise to the conversations to make them more interesting, because without noise the conversations will go entirely as planned. With the added noise, agents might receive the wrong signal or no signal at all, which increases the likelihood of overlapping speech and longer silences. Figure 4.1 shows an example fragment of a conversation that was generated with the Conversation Simulator, including the signals that were sent.

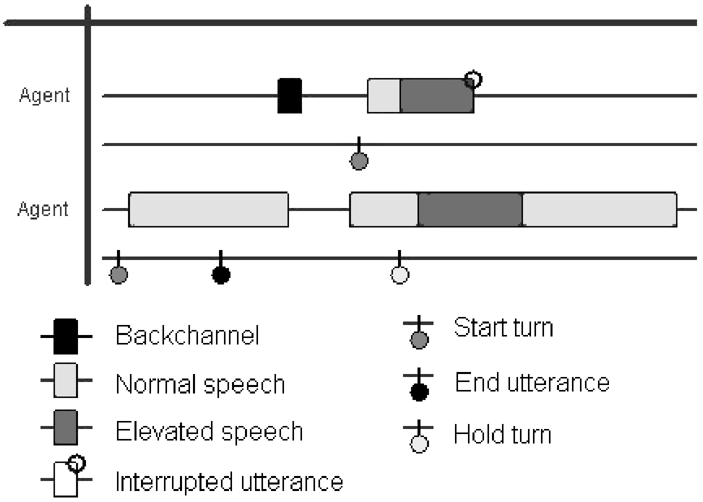

Figure 4.1: An example fragment of a conversation as generated by the Conversation Simulator. The image was taken from Ter Maat and Heylen (2010).

The behaviour of the two agents can be scripted. In these scripts one can define how an agent reacts to different situations. The core conversation simulator runs the scripts of the agents in parallel and takes care that the rules are executed and the variables are updated accordingly. An example could be:

```
<rule name="start__after_detect_endOfTurn">
 <precondition name="own_state" value="silence" />
 <precondition name="compare" value1="time_in_own_state"
               comparator="greater_than" value2="200" />
 <precondition name="detected_signal" value="end_of_turn" />
 <reaction action="change_parameter" name="#want_start" value="1" />
</rule>
```

The agent that has this rule first checks if it has currently been silent for longer than 200 ms (line 2 and 3), and if it was just received an END_OF_TURN signal (line 4). If these preconditions are fulfilled then the parameter #WANT_START is set to one (line 5), which causes the agent to start speaking.

The conversations of the agents can be visualized (Figure 4.1) and made audible. The visualization is drawn real-time and saved as an image at the end of the conversation. For the speech rendition, we wanted to output natural but *incomprehensible* speech, for two reasons. First, for an automatically generated conversation it is easier to take incomprehensible speech, instead of saying something meaningful. The latter obliges the system to pay attention to the coherence of the conversation's content, which is a lot harder and not the focus of this study. Secondly, we wanted to study how different turn-taking strategies affect the perception of the agents. Having comprehensible speech adds a lot of variables to the study, because the content of the speech can severely influence this perception. To create this incomprehensible speech, we extracted several sentences with a clear start and end point from the AMI corpus[1]. These sentences were fed through a Pass Hann Band Filter, from 0Hz to 500Hz with a smoothing of 100Hz. With this method the sentences kept their prosodic information but lost their content. When an agent starts to say something in the simulation, a random sentence is selected and played until the end or until interrupted.

4.2 Turn-taking strategies

The procedures for turn management that we considered in this study consisted of *start-up* strategies and *overlap resolution* strategies. For each group we defined three possible strategies, yielding nine different turn-taking strategies in total by crossing them.

The start-up strategy determined when to start a new utterance — that is, a speech turn. We considered the following three strategies:

1. EARLY: a speaker will start their turn just before the end of the interlocutor's turn

2. DIRECT: a speaker will start their turn immediately after the interlocutor's turn has finished

3. LATE: a speaker will leave a pause (of a few seconds) before he or she starts their turn after the interlocutor's turn has finished.

The overlap resolution strategy determines how to behave when there is overlapping speech — that is, when two persons are speaking at the same time. We considered the following three strategies:

1. STOP: a speaker will decide to stop speaking

2. NORMAL: a speaker will decide to continue speaking normally

3. RAISED: a speaker will decide to continue speaking with a raised voice

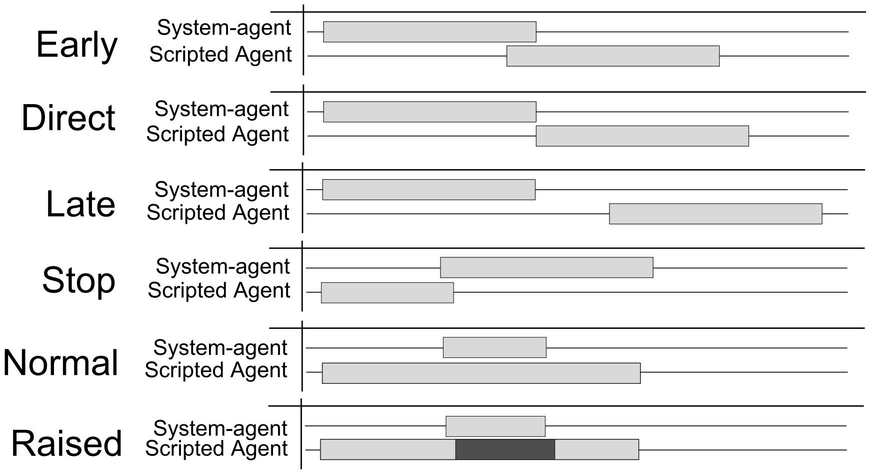

Figure 4.2: A graphical representation of the turn-taking strategies that were used. The dark section represents a raised voice.

Figure 4.2 shows a graphical representation of the different turn-taking strategies.

Of course in everyday conversations, people use a mixture of these strategies for different circumstances, but since the goal is to study the effects of a single turn-taking strategy, only a single strategy was used in each conversation. The strategies that were selected were based on a study by Schegloff (2000).

Using the conversation simulator, eight different agents were scripted using the turn-taking strategies described, namely all combinations of start-up strategy and overlap resolution strategy, minus the strategy EARLY+STOP. This strategy was discarded because it would start while the other person was speaking and immediately stop again, resulting in very awkward conversations. The other eight scripts resulted in very different conversations. In Figure 4.3, two examples of different conversations are shown. The contributions of the agent that varies its start-up and overlap resolution strategy (which we will refer to as the scripted agent) are shown on the lower tier. The fixed system-agent (which we will refer to as the system) was programmed to use different strategies based on chance and is shown on the top tier in each case. As the system-agent uses a random turn-taking strategy *each turn*, all scripted agents encounter all possible turn-taking strategies during a conversation, which makes the comparison of the agents more fair. Note that the conversations in Figure 4.3 are quite different. The question now is whether these interactions lead to different perceptions.

4.3 Experimental Setup

The variation in the turn management scripts results in different interaction patterns which might change the impression of how the agents relate to each other on an interpersonal scale, or they it change the impression of the personality of the agent.

In our study we invited ten people to rate the eight conversations on 13 dimen-

[1] http://www.amiproject.org

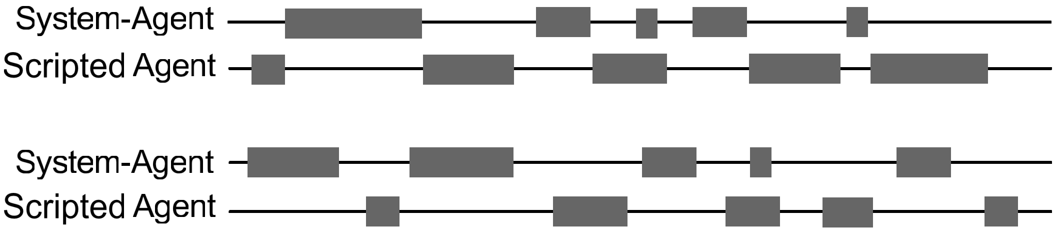

Figure 4.3: Example conversations with different turn-taking strategies. Top pane: DIRECT+RAISED, bottom pane: LATE+CONTINUE.

sions. The raters were all between 20 and 30 years old, mainly students, 6 male and 4 female. We asked them to rate the agent on semantic differential scales: pairs of bipolar adjectives places at the extremes of 5-point Likert scales. The adjectives were taken from Fukayama et al. (2002) and Goldberg (1993), and because of the nature of the various SEMAINE characters (Mcrorie et al., 2009) we also added some adjectives that are related to the different characters. Table 4.1 shows all semantic adjectives that were used and their sources. The reason we selected these adjectives is because they are useful for the SEMAINE characters or because we think turn-taking affects them.

Unfriendly - Friendly^a	Disagreeable - Agreeable^b
Cold - Warm^a	Passive - Active^c
Undependable - Responsible^a	Negative - Positive^c
Rude - Respectful^a	Not aroused - Aroused^c
Distant - Pleasant^a	Unattentive - Attentive^c
Unpredictable - Stable^b	Submissive - Dominant^c
Negligent - Conscientious^b	

Table 4.1: Semantic differential adjectives used in questionnaire. The superscripts indicate that the adjectives were selected based on (a) Fukayama et al. (2002), (b) Goldberg (1993), or (c) the SEMAINE characters.

The raters were seated in front of a PC which ran a Powerpoint presentation. On each slide they could click on an audio file that would then play. The audio of the system agent came from the left loudspeaker and the audio from the scripted agent which they had to rate from the right loudspeaker. We made sure that each rater knew which speaker they had to rate. To make the difference even clearer, the scripted agent's speech was somewhat higher in pitch than that of the system agent. It could be that a higher pitch leads to changes in the perception, but as in every recording the agent's voice had the same (higher) pitch the effects were the same for each recording, and therefore the same for each turn-taking strategy.

The conversations were ordered such that conversations in which the system was more talkative than the scripted agent alternated with conversations in which the scripted agent was more talkative. We had five raters listen to this order (A) and five raters listened to an order in which the first four conversations of A changed position with the last four conversations of A. After each conversation the raters were asked to complete the questionnaire on how they perceived the agent whose speech came from the right loudspeaker.

4.4 Results

In this section, we will show the results of the PASSIVE experiments. We show the average ratings of the dimensions of the different turn-taking strategies and their significant differences. After that, we will try to reduce the number of dimensions by performing a factor analysis.

4.4.1 Rating results

The mean ratings and their standard deviations of the different strategy-combinations for each dimension can be found in Table 4.2. Using boldface we have marked the conversation that received the highest mean rate for a scale and using underlinings we have marked the conversation with the lowest mean rate for the scale.

Dimension	Early+Normal		Late+Stop		Early+Raised		Direct+Stop		Direct+Normal		Late+Raised		Direct+Raised		Late+Normal	
	Avg	Sd	Avg	Sd	Avg	Sd	Avg	Sd	Avg	Sd	Avg	Sd	Avg	Sd	Avg	Sd
Negative - Positive	2,6	1,1	2,6	0,5	2,1	0,9	**3,7**	0,7	2,8	0,9	2,4	0,5	2,3	0,7	3,1	0,9
Not aroused - Aroused	3,0	1,2	3,1	1,1	**4,8**	0,4	3,3	0,9	3,9	0,6	3,7	1,1	4,1	0,7	3,0	1,1
Unfriendly - Friendly	3,3	1,1	3,0	0,0	2,2	0,9	**3,8**	0,8	3,5	1,0	3,4	0,5	3,1	1,3	3,7	0,7
Disagreeable - Agreeable	2,4	1,2	3,1	1,5	1,7	0,8	**3,5**	1,1	2,3	0,7	2,2	0,6	2,4	1,1	3,3	0,8
Negligent - Conscientious	3,9	1,2	3,0	1,2	3,2	1,2	3,7	0,8	3,7	1,1	3,1	0,9	**4,0**	0,7	3,4	0,5
Rude - Respectful	3,2	1,2	3,7	0,8	2,0	1,2	**4,2**	0,8	3,0	1,1	2,9	0,7	2,9	0,9	4,1	0,6
Distant - Pleasant	2,6	0,8	2,5	0,7	2,5	0,8	**4,2**	0,8	2,7	0,8	2,6	1,0	3,0	1,1	3,1	1,0
Unpredictable - Stable	3,7	0,8	3,3	1,2	2,0	0,7	3,6	1,1	3,4	1,0	2,7	1,3	2,8	1,0	**3,9**	0,7
Unattentive - Attentive	3,4	1,2	3,3	0,5	3,0	1,3	**4,3**	0,5	3,5	1,1	3,5	0,7	3,7	0,9	3,6	0,5
Cold - Warm	2,5	1,2	2,9	0,6	2,4	0,5	**3,9**	0,6	2,9	0,9	2,7	0,7	2,7	0,9	3,2	0,8
Passive - Active	4,4	0,5	2,6	1,5	4,7	0,5	3,5	0,5	**4,8**	0,4	3,5	1,0	4,0	0,7	2,8	0,8
Submissive - Dominant	4,4	1,3	1,1	0,3	4,3	0,8	3,3	0,8	**4,5**	0,7	3,3	1,1	3,7	1,3	2,7	0,5
Undependable - Responsible	3,2	1,0	2,9	0,7	2,7	1,2	**3,7**	0,8	3,4	0,7	2,4	0,8	3,3	0,9	3,3	0,7

Table 4.2: Results passive experiment: average scores of 5-point semantic differential scales. Underlined values are the lowest of that row, and bold values are the highest.

First, when one looks at the bold face figures it is immediately obvious that the DIRECT+STOP column — in which the agent starts speaking immediately when the system stops and stops speaking when overlap is detected — attracts most of the high scores. It is the top highest rated version on positivity, friendliness, agreeability, respect, pleasantness, attentiveness, warmth and responsibility. On the other hand, EARLY+RAISED — where the agent starts before the system has ended and raises its voice in the case of an overlap — is the highest rated version on negativity, unfriendliness, disagreeability, rudeness, distance, unpredictability, unattentiveness, and coldness. It was also rated as the most "aroused" agent. These two strategies appear to be the most extreme.

What to think about the other highest and lowest scores? The EARLY+NORMAL and LATE+RAISED agents show the lowest scores on the arousal and responsibility dimensions respectively. Interestingly though, the lowest score on arousal is 3.0 (which is shared between Early+Continue and Later+Normal), which means that there is no agent that effectively performs low on arousal. Similarly, the current strategies do not yield an agent that is extremely negligent. The opposite, conscientiousness was rated highest when an agent did not wait (long) before the turn of the system has finished and continues (possibly speaking louder) when overlap was detected; that is, the Early+Normal and Direct+Raised strategies scored highest on conscientiousness. The impression of highest stability is reserved for the agent that started speaking only after the system had finished but did not mind continuing when interrupted (without raising its voice), Late+Normal. An agent that waited before the other had finished but raised its voice when interrupted, Late+Raised, on the other hand, was not considered to be dependable.

Dimension	Start-up Strategy						Overlap Resolution Strategy					
	Early		Direct		Late		Stop		Normal		Raised	
	Avg	Sd	Avg	Sd	Avg	Sd	Avg	Sd	Avg	Sd	Avg	Sd
Negative - Positive	2,4	1,0	2,9	0,9	2,7	0,7	3,1	0,8	2,8	0,9	2,3*	0,7
Not aroused - Aroused	3,9	1,3	3,8	0,8	3,3	1,1	3,2	1,0	3,3	1,0	4,2***	0,9
Unfriendly - Friendly	2,8*	1,1	3,5	1,0	3,4	0,6	3,4	0,7	3,5	0,9	2,9*	1,1
Disagreeable - Agreeable	2,1*	1,1	2,7	1,1	2,9	1,1	3,3	1,3	2,7	1,0	2,1*	0,9
Negligent - Conscientious	3,6	1,2	3,8	0,9	3,2	0,9	3,4	1,1	3,7	1,0	3,4	1,0
Rude - Respectful	2,6*	1,3	3,4	1,1	3,6	0,9	4,0	0,8	3,4	1,1	2,6*	1,0
Distant - Pleasant	2,6	0,8	3,3	1,1	2,7	0,9	3,4	1,1	2,8	0,9	2,7	1,0
Unpredictable - Stable	2,9	1,1	3,3	1,0	3,3	1,1	3,5	1,1	3,7	0,8	2,5*	1,0
Unattentive - Attentive	3,2	1,2	3,8	0,9	3,5	0,6	3,8	0,7	3,5	0,9	3,4	1,0
Cold - Warm	2,5*	0,9	3,2	0,9	2,9	0,7	3,4*	0,8	2,9	1,0	2,6	0,7
Passive - Active	4,6*	0,5	4,1***	0,8	3,0*	1,2	3,1**	1,2	4,0	1,1	4,1	0,9
Submissive - Dominant	4,4	1,0	3,8	1,1	2,4***	1,2	2,2***	1,3	3,9	1,2	3,8	1,1
Undependable - Responsible	3,0	1,1	3,5	0,8	2,9	0,8	3,3	0,9	3,3	0,8	2,8	1,0

Table 4.3: Results passive experiment: average scores of 5-point semantic differential scales, * = p <0.05, ** = p <0.01, *** = p <0.001

Table 4.3 contains the fragments grouped by start-up strategy and overlap resolution strategy, with their mean and standard deviation. The significance was calculated for every dimension scale by performing a two-paired t-test for all combinations (1+2, 2+3, 1+3). The type of t-test (equal variance or unequal variance) was determined by performing an f-test first. A strategy was said to be significantly different when both t-tests with the other strategies scored a $p < 0.05$. So, for example, the value of the dimension 'negative' for raising the voice was significantly different because both the t-test with stop and the t-test with continue resulted in a $p < 0.05$. The left part of figure 4.3 shows that, for the start-up strategy, most significant differences occurred with the situation in which the agent started before the system was finished. Starting EARLY was seen as more unfriendly, disagreeable, rude, cold and more active, compared to starting DIRECTLY or starting LATE. The most pleasant person would be a speaker who started directly at the end of the other person's speech, not sooner or later. Also, the sooner a person started to talk the more active he or she was perceived: the Early strategy was perceived significantly most active, and the Late strategy was perceived significantly most passive. A final (highly) significant result is that the agent with the Late strategy was perceived as most submissive.

The right-hand part of Table 4.3 shows the mean ratings and standard deviations of the results grouped by overlap resolution strategy. It shows that especially the

Stop and Raised strategy were significantly different on several dimensions. Stopping when the agent detects an overlap was perceived as warmer, more passive and more submissive than continuing (whether normal or with a raised voice). Continuing with a raised voice was perceived as more negative, more aroused, less friendly, less agreeable, more rude and more unpredictable than stopping or continuing normally.

4.4.2 Grouping results

Because we suspected that certain dimensions correlated with each other, we decided to create a better overview of the results by reducing the number of scales. Using a factor analysis (Principal Component Analysis with a Varimax rotation with Kaiser normalization), we grouped the items with a correlation > 0.5 and created four factors (see Table 4.4).

Factor	Low value		High value	Correlation
1. *agreeableness* $\alpha = 0.886$	Distant	-	Pleasant	0.85
	Cold	-	Warm	0.81
	Negative	-	Positive	0.78
	Unfriendly	-	Friendly	0.76
	Disagreeable	-	Agreeable	0.63
	Rude	-	Respectful	0.58
2. *assertiveness* $\alpha = 0.766$	Submissive	-	Dominant	0.90
	Passive	-	Active	0.84
3. *neuroticism* $\alpha = 0.703$	Aroused	-	Not Aroused	0.81
	Unpredictable	-	Stable	0.78
4. *conversational skill* $\alpha = 0.679$	Negligent	-	Conscientious	0.87
	Unattentive	-	Attentive	0.72
	Undependable	-	Responsible	0.63

Table 4.4: Results of the factor analysis applied to all scores of the scales (factors and scales ordered by Cronbachs alpha value and correlation strength)

The five main personality traits, such as reported by Goldberg (1993), can be used to describe the first three factors. Factor 1 can be described by *agreeableness*, with high values corresponding with someone who is cooperative, compassionate, friendly and optimistic. Factor 2 is similar to the trait *assertiveness* (previously used by Robinson and Reis (1989)), and is associated with someone who is extravert, strong, and pushy. Factor 3 is best covered with the trait *neuroticism*, and is associated with people who are more calm and emotionally stable. Factor 4 can best be described with the term *conversational skill*, meaning how adept a person is in a conversation.

Now we can calculate the average values for the different turn-taking strategies and significances. Figure 4.4 shows the values of the four factors for the three start-up strategies, and Figure 4.5 shows the values of the four factors for the three overlap resolution strategies.

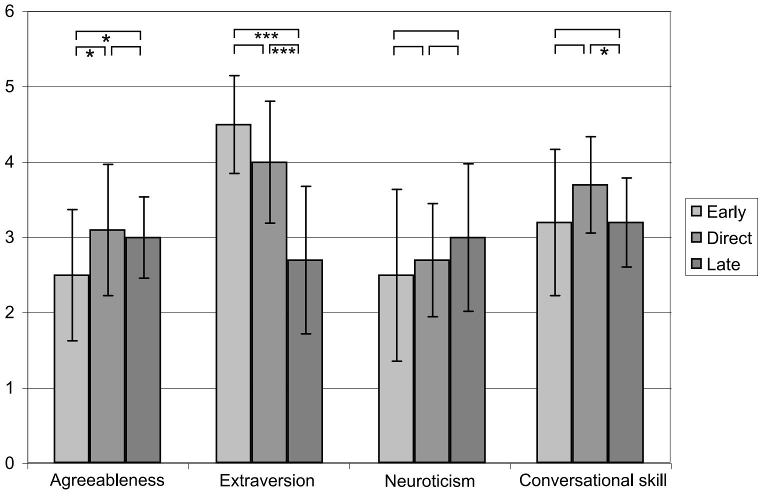

Figure 4.4: The results of different start-up strategies. * = p <0.05, ** = p <0.01, *** = p <0.001

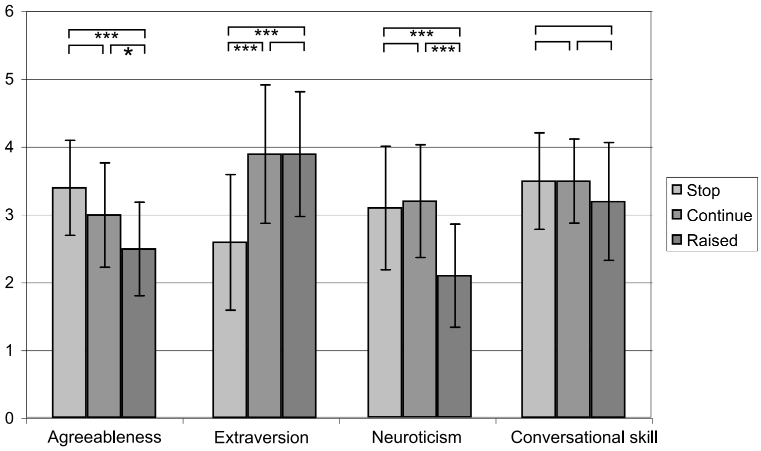

Figure 4.5: The results of different overlap resolution strategies. * = p <0.05, ** = p <0.01, *** = p <0.001

These figures show that for start-up strategy, starting EARLY is seen as significantly less agreeable and more assertive, and starting LATE the opposite. Interestingly, starting DIRECT is perceived as having more conversational skill. For the overlap resolution strategy, STOPPING is perceived as significantly more agreeable and less extravert. In contrast, continuing with a RAISED voice is perceived as less agreeable, more extravert, and less neurotic (more active and unpredictable).

4.5 Summary

In this chapter, we described our first study on the perception of turn-taking strategies. We studied three different start-up strategies — in which the participants started just before the other person's turn-end, directly when that turn ends, or after a certain pause — and three different overlap resolution strategies — in which the participant stops speaking, continues normally or continues with a raised voice after overlapping speech is detected. Using a program that can simulate conversations, we programmed one of the agents to follow the studied turn-taking strategies and let it 'talk' with another agent that used random turn-taking behaviour. This resulted in conversations containing non-intelligible speech, which were given to human raters. These raters had to write down how they perceived the agent that followed the studied turn-taking strategies using 13 semantic differential scales.

We found that starting before the end of the other person's turn was perceived as more extrovert and less agreeable, while starting after a small pause was perceived as more introvert and less agreeable. Stopping when overlapping speech is detected was perceived as more agreeable and introvert, while continuing with a raised voice was perceived as less agreeable, more extrovert and neurotic.

In the next chapter, we will present our second study, in which humans actively participated in a conversation in which the interviewing agent (controlled by a wizard) used different turn-taking strategies to start its questions.

5 Turn-taking perception when interacting with an interviewer

In the previous chapter, the perception of different turn-taking strategies was studied by generating incomprehensible conversations, in which one interlocutor used different turn-taking strategies. Human raters had to listen to these conversations and, using a questionnaire, write down how they perceived that interlocutor. This meant that the rater was an observer of the conversation, merely a bystander who did not actively participate. By actively involving the participant in the conversation, we think that the perception of the agent's behaviour would be much stronger, if not different. For example, one can imagine that hearing person A interrupting person B would be perceived differently from being interrupted oneself, because it is much more annoying if you are the one who is not allowed to finish your turn.

In this chapter we present a second experiment in which, in contrast to the PASSIVE experiment in the previous chapter, the subjects were not passive bystanders anymore (Ter Maat et al., 2010). In the ACTIVE Wizard of Oz experiment, the subjects themselves participated in the human-agent interaction, acting as interviewees. The interviewer was an agent controlled by a human wizard who determined which turn-taking strategy was applied by the agent. Participants conversed with this interviewing agent and completed a questionnaire afterwards. Similar to the PASSIVE experiment, we aimed to investigate how agents' turn-taking strategies influenced users' impressions of these agents. Additionally, we wanted to know how these turn-taking strategies influenced the users' response behaviour.

We adopted the interviewing domain for several reasons. The first reason is that it allows for some control over turn-taking behaviour and conversations. In this setting, the responses of the agent are known beforehand (as a list of questions) and are not changed during the conversation. During the conversation, the agent knows what its next question is, and the only thing that has to be determined is when to start asking it — independent of what the answer of the participant is. The second reason is that interviews have a conversational style similar to that of Sal: both conversations are unbalanced, that is, one person is more important and speaking more often than the other person. The goal is to have one person (the speaker in SEMAINE or the interviewee) keep on talking. In the SAL domain, this goal is reached by asking for more information or changing the topic by asking a question about another subject. In the interviewing domain, this goal is reached by continuously asking questions.

Of course there are also some downsides to using the interview domain. First of all, limiting the study to only one domain might cause problems with the generalizability of the results. Behaviours can have different impacts and meanings in different types of conversations. Secondly, following this point, the choice of the interviewing domain influences the results. Interviews are, in general, more polite and have more social rules than say a conversation between two friends. This means, for example, that interrupting the interviewee is probably perceived as more rude, while letting the interviewee finish their story is perceived as more polite. We argue, because interviews are polite by nature, that the results of this study can at least be generalized to other polite conversations and settings, such as a virtual receptionist or guide. We have to mention though that we do not have any evidence for this claim, and leave this question open for further research.

The turn-taking strategies that we used in the ACTIVE experiment are almost similar to the start-up strategies EARLY, DIRECT, and LATE, described in section 4.2. In this case, however, the overlap resolution strategies were not taken into consideration because during a pilot test, we noticed that overlapping speech was always avoided by the human participant. This means that the overlap resolution strategies could only marginally come into effect. Note that this does not say anything about whether the overlapping speech was 'correct' or not, because the human subject could also stop speaking because he or she was almost finished. Another important difference with the PASSIVE experiment is that in the ACTIVE study, the strategies were applied by a human wizard rather than by a computer agent that was programmed to start EARLY, LATE, or DIRECT. As humans are not machines this meant, in practice, that the human wizard would start its turns (i.e., questions) at variable time intervals before or after the user's turn, and that not all turns were started at the exact same times with respect to the strategy. Hence, prior to the analysis of the subjects' ratings, we analysed the speech recordings of the experiment and verified that the timing of the questions was in accordance with the intended strategy.

Because of the polite nature of interviews, we expected that an interviewee would rate an agent more polite and positive if he or she had more time to answer.

5.1 Experimental Set-Up

In this section we describe how the interview sessions were set up and arranged, and what type of topics and questions were used for the interviews. Furthermore, we describe the semantic differential scales used in the questionnaire.

5.1.1 Participants

Twenty-two people participated in the experiment, most of them were PhD students. There were sixteen male and six female subjects with an average age of 27.55 (standard deviation of 3.41). Each of them signed a form, giving consent to the use of recordings of their speech for research purposes.

5.1.2 Stimuli: scenarios of interviews

In contrast to the PASSIVE study, in this study the conversations did have a certain content and context. It should be noted that the interview setting with the agent in the role of an interviewer constrains the flow of the conversation as the initiative lies mainly with the agent. This allows us to limit the number of utterances the agent should be able to say. The agent asks a question, and independent of the actual answer of the user, the agent (or Wizard rather) anticipates the user's turn-end and then asks the next question using one of the three start-up strategies.

In such a setup, the agent's questions are very important. We designed the questions such that they would be easy to answer since a complex question, which requires more thinking, can disrupt the flow of the conversation. Also, the questions asked by the agent were designed so as not to be answerable with one or two words only, since it is hard to apply a certain start-up strategy when the user only says 'Yes'. Examples of the questions used are "Can you tell me what movies you like?", and "Which restaurants do you recommend and why?".

Another possible problem is that certain questions can influence the results because each question has certain connotations that are perceived differently by each user. For example, person A could really like a certain topic, while person B absolutely hates it. Therefore, to improve the generalizability, we decided to create three sets of questions, each on a different topic, namely 'food and drinks', 'media' and 'school and study'. By making three different groups it was possible to interchange the questions used in each session (a single conversation of a user with the system). This decreases the influence of the questions on the results. Also, by making three sets of related questions, the questions fit in the same context and do not disrupt the flow of the conversation.

Another factor to consider is the voice to use. The difference between a male or a female voice can greatly influence the perception of the user. One voice may sound more friendly than another, or male and female participants may listen differently to male and female voices. To control for this variable we introduced two agents: one with a male and another with a female voice. These voices were changed each session to decrease the influence of the voice on the results. The spoken questions were synthezised with the Loquendo TTS software[1].

Each session (i.e., a single conversation of the user with the agent) therefore followed a scenario that consisted of a certain start-up strategy (EARLY, DIRECT, or LATE), a certain voice (MALE or FEMALE), and a certain topic ('food and drinks', 'media', and 'school and study'). The exact properties — start-up strategy, voice and topic — of each session were randomized and counterbalanced.

5.1.3 Recordings

Speech recordings (mono, 44.1 kHz) were made of each session with a microphone that was placed on a desk near the subject. The microphone captured both the voices of the interviewing agent and the subject. The speech recordings were made to 1)

[1] http://www.loquendo.com

validate the wizard-enforced turn-taking strategies, and 2) to investigate the effects of turn-taking strategies on the subject's speech behaviour.

5.1.4 Procedure

The participants were told that they would talk with a speech-to-speech dialogue system, with the agent in the role of an interviewer. They were told that we had implemented different personalities in different parts of the dialogue system, and that their role was to write down, in a questionnaire, how they perceived each agent. After this introduction they talked with the agent three times (three sessions), each session with a different scenario. These scenarios were created in such a way that every possible combination and order was used at least once.

During each session the participant sat in front of a microphone and a set of loudspeakers. The wizard sat behind the participant, out of sight but in the same room. During the interview, the wizard would follow the current start-up strategy by clicking on the button to start the next question at the right time (as intended by the strategy).

Although the wizard in this study was not blind — that is, he was aware of the aims of the study — we expected and presumed this would not affect the results of this study. The only aspect that the wizard could control was the start time of the next question, and a manipulation check (see section 5.2) showed that the human wizard was able to apply the strategies correctly. Our analyses were carried out on the users' responses which were only influenced by the starting times of the questions.

5.1.5 Measures: questionnaire design

After the interview, the subjects completed a questionnaire about how they perceived the interviewer using 7-point Likert scales (see section 4.3 and Table 5.1) to capture the perceived impressions of the users.

Disengaged - Engaged	Competitive - Cooperative
Aggressive - Calm	Impolite - Polite
Closed - Open	Introvert - Extravert
Weak - Strong	Inexperienced - Experienced
Pushy - Laidback	Shy - Bold
Arrogant - Modest	Insecure - Confident
Not socially skilled - Socially skilled	Tensed - Relaxed
Distant - Close[2]	Careless[3] - Responsible

Table 5.1: Semantic differential adjectives used in the ACTIVE experiment, as an addition to the original set in section 4.3 that was used as well.

[2] This was called 'Pleasant' in the PASSIVE experiment.

[3] This was called 'Undependable' in the PASSIVE experiment.

We started with the same semantic differential scales as presented in section 4.3, and extended these with new scales that were tuned to the specific interviewing setting, to capture more social-skills-related attributes and the interviewer' interviewing capabilities. Table 5.1 shows the additional dimensions. For two scales that were adopted from the PASSIVE experiment, the adjectives were changed because we felt that these adjectives would better describe the interpersonal attitude that we intended to measure.

5.2 Manipulation Check

Since applying the correct strategy by a human wizard is error-prone, one requires an objective measure to see how consistently each start-up strategy was applied. We therefore annotated the speech recordings of the interviews on who was speaking when and we looked at the two objective measures *gap length* and *number of overlaps*. The gap length is the duration of silence between the end of the user's turn and the start of the agent's next question, see Fig. 5.1. The gap length should be shortest for the EARLY strategy and longest for the LATE strategy, and should be significantly different from each other for each strategy. The number of overlaps is the average number of overlaps per session: an overlap occurs when the agent starts the next question while the user is still speaking, see Fig. 5.2. The number of overlaps should be highest for the *Early* strategy, and lowest for the *Late* strategy.

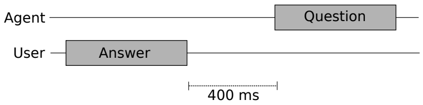

Figure 5.1: An example of a gap with a duration of 400 ms

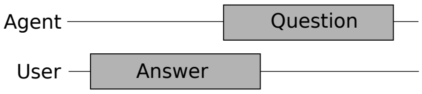

Figure 5.2: An example of an instance of overlap

Table 5.2 shows the average gap length between the user's current turn and the following interviewer's question, grouped by the start-up strategy that was used. As we can observe, the gap length is shortest for the EARLY strategy and longest for the LATE strategy. The differences between the gap lengths of the 3 strategies are highly significant ($p < 0.001$).

One might expect at first that the EARLY strategy would not result in silences between the turns, since the goal of this strategy was to respond *before* the other person was finished speaking. However, sometimes the wizard's predictions of the end-of-turn were off, and the wizard responds too late, leaving a small gap between the turns. This is acceptable as long as the average duration of the silence is shorter than for the other two strategies, and the total number of overlaps is larger than for the other two strategies. These requirements were both met, as shown by table 5.2 and Figure 5.3.

Figure 5.3 shows the average number of overlaps grouped by the start-up strategy that was used. As shown, the number of overlaps is highest in the EARLY strategy and lowest in the LATE strategy. The difference between the EARLY strategy and the other two strategies is highly significant ($p < 0.001$), but the difference between the DIRECT

Strategy	Gap length	
	Mean	Sd
EARLY	0.72	0.69
DIRECT	1.07	0.58
LATE	1.97	0.57

Table 5.2: Gap lengths between the user's turn and the following interviewer's question

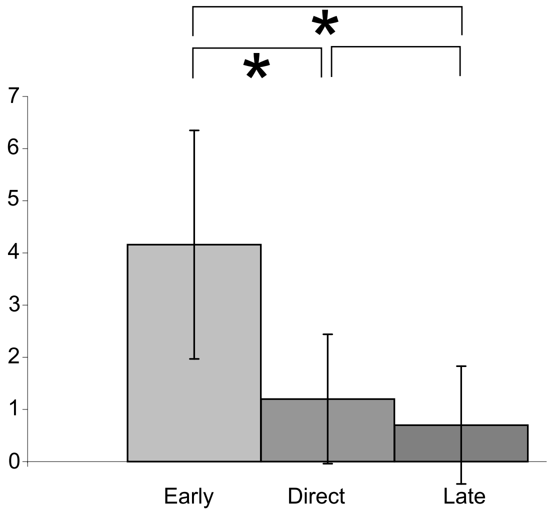

Figure 5.3: Number of overlaps (user and interviewer speaking at the same time. * = $p < 0.05$

and the LATE strategy is not. Because both the direct and the late strategy wait for the end of the user's turn, we did not expect any significant difference between these two strategies for overlaps.

These results show that there was indeed a significant difference between the start-up strategies in accordance with the desired effect, which means that the turn-taking strategies were correctly applied.

5.3 Results

Here, we present the results of the ACTIVE experiment. First, the number of semantic differential scales used in the questionnaire is reduced by applying a factor analysis that groups scales into a smaller number of so-called factors. Subsequently, using these factors, we will analyse the users' ratings with respect to the various turn-taking strategies to see the effect of the turn-taking strategy on the user's impressions. In addition, we will look at how the turn-taking strategies influence the user's response behaviour by analysing the user's speech from the recordings.

Factor	Low value		High value	Correlation
Factor_1 (*agreeableness*) $\alpha = 0.885$	Cold	-	Warm	0.86
	Unfriendly	-	Friendly	0.78
	Tensed	-	Relaxed	0.72
	Disagreeable	-	Agreeable	0.70
	Aggressive	-	Calm	0.63
	Competitive	-	Cooperative	0.60
	Negative	-	Positive	0.60
	Impolite	-	Polite	0.52
Factor_2 (*assertiveness*) $\alpha = 0.878$	Strong	-	Weak	0.85
	Dominant	-	Submissive	0.79
	Extravert	-	Introvert	0.76
	Bold	-	Shy	0.73
	Arrogant	-	Modest	0.57
	Pushy	-	Laid back	0.53
Factor_3 (*conversational skill*) $\alpha = 0.849$	Inexperienced	-	Experienced	0.72
	Not socially skilled	-	Socially skilled	0.72
	Unpredictable	-	Stable	0.69
	Careless	-	Responsible	0.60
	Unattentive	-	Attentive	0.50
Factor_4 (*rapport*) $\alpha = 0.833$	Closed	-	Open	0.82
	Disengaged	-	Engaged	0.71
	Distant	-	Close	0.62
	Negligent	-	Conscientious	0.58

Table 5.3: Results of the factor analysis applied to all scores of the scales (factors and scales ordered by Cronbachs alpha value and correlation strength)

5.3.1 Grouping scales in the questionnaire by factor analysis

As a way to reduce the number of scales (we used 27 scales, see section 4.3 and 5.1), a factor analysis was performed to see whether some of the scales could be grouped together. We used a Principal Component Analysis with a Varimax rotation with Kaiser normalization. From the results we grouped the items with a correlation > 0.5, which resulted in four different factors. These four factors, the corresponding scales and the corresponding correlations can be found in Table 5.3.

Intuitively, the grouping of the scales makes sense. Factor 1 can be described as *agreeableness*, one of the five main personality traits as reported by Goldberg (1993). A high level of agreeableness corresponds to someone who is cooperative, compassionate, friendly, and optimistic. Next, the adjectives strong, dominant, extravert, bold, arrogant, and pushy, can be covered under the term *assertiveness* (which has previously been used by Robinson and Reis (1989) in a similar context). The third factor contains scales that have more to do with *conversational skills* of the agent (the interviewer). The last factor seems to be related to *rapport* (Gratch et al., 2006). High rapport means that the participants are 'in sync' or 'on the same wavelength', which in turn means that the participants are very close and engaged during the interaction. There are four scales which show too low correlations with any of the factors, namely: RUDE-RESPECTFUL, NOT AROUSED-AROUSED, INSECURE-CONFIDENT, and PASSIVE-ACTIVE.

5.3.2 Analysis of subjects' ratings

In order to see the effects of the strategies on the ratings in the factors and scales, an ANOVA test was performed on the data with Bonferroni correction. We used the ratings from the four factors found in the previous section, and the ratings from the four scales that did not fit in these factors.

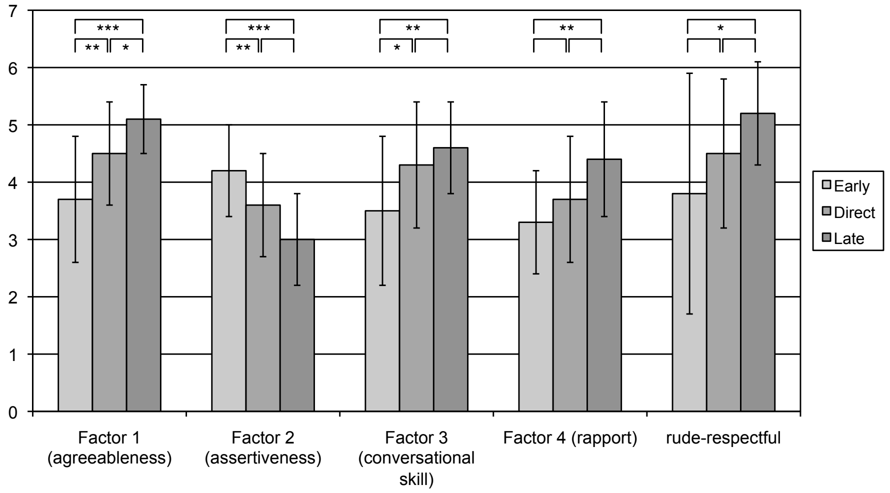

Figure 5.4: The results of the different start-up strategies. * = p <0.05, ** = p <0.01, *** = p <0.005

Figure 5.4 shows the results of the four factors and the rude-respectful scale (the other three scales that did not fit the factor analysis did not provide any significant results). All factors indicate a significant difference between the EARLY and LATE strategy. The strongest factor is Factor 1, *agreeableness*, where the ratings for all three strategies are significantly different ($F(2, 66) = 17.06, p < 0.001$). Starting EARLY was seen as more unfriendly, tensed, aggressive, competitive and negative, and starting LATE was perceived as more friendly, relaxed, agreeable, cooperative and positive. For the factor *assertiveness*, the EARLY strategy differs significantly from the DIRECT and LATE strategy ($F(2, 66) = 12.48, p < 0.001$), but there is no significant difference between DIRECT and LATE. Starting EARLY was rated as more strong, dominant, extravert and bold. Similarly, for *conversational skill*, the DIRECT and LATE strategies do not differ significantly from each other, but the EARLY and LATE strategy do ($F(2, 66) = 6.24, p = 0.003$). People perceived agents starting EARLY as more inexperienced, less socially skilled, more unpredictable, more careless and unattentive than agents who start LATE. *Rapport* was significantly differently perceived in the EARLY and LATE strategy ($F(2,66) = 6.52, p = 0.003$), with starting LATE having significantly more perceived rapport than starting EARLY. Finally, starting late was perceived significantly more respectful than starting early ($F(2,66) = 4.33, p = 0.18$).

5.3.3 Analysis of the subjects' speech behaviour

The previous section shows that different turn-taking strategies change the perception that the user has about the agent. In this section we will try to find out whether these different turn-taking strategies affected the behaviour of the user as well. Under the assumption that people, in general, accommodate their speech behaviour to the speech behaviour of their conversational partner (Giles et al., 1973; Staum Casasanto et al., 2010), it makes sense to expect that certain turn-taking strategies will influence the subjects' speech behaviour. Listening to the recordings of the interviews, we observed that some of the subjects started to speak faster and make shorter turns when the EARLY strategy was applied. Intuitively, this observation makes sense, in that people may try to avoid interruptions by speaking faster and by making shorter turns in order to finish those turns earlier. Hence, we measured speech rate and the lengths of the subjects' turns (i.e., their answers) and analysed these with respect to the turn-taking strategy applied.

Speech rate In order to measure speech rate, we first extracted all the user turns from the recordings. Using Praat (Boersma and Weenink, 2001) and a script that automatically detects syllable nuclei (De Jong and Wempe, 2009) we extracted the speech rates of all turns, where we define speech rate as the number of syllables per second. To compensate for people having different speech rates, we normalized the speech rates for each user by subtracting the average speech rate of that user and dividing the result with the standard deviation of that user's speech rates.

Using these speech rates, we want to verify whether people (consciously or unconsciously) changed their behaviour when they noticed that the agent used a certain turn-taking strategy. We assume that it took some time for people to adjust to the interviewer's behaviour so we split all the conversations into two parts, and compared the normalized speech rates of the first half with the speech rates of the second half to see whether accommodation took place. Table 5.4 contains the results of this comparison.

Table 5.4 shows that especially in the EARLY strategy people changed their speech rates. In the second half of the conversations with an agent using the EARLY strategy, users increased their speech rate compared to the first half, probably because they wanted to finish their sentence before the agent could interrupt them. So, only in the EARLY strategy did the subjects accommodate to the interviewer's behaviour by speaking faster.

Next, we compared the speech rates found in the different strategies with each other to find out whether a subject spoke faster in one strategy than another. The results of this can be found in Table 5.5.

Table 5.5 shows that in the first part of the EARLY strategy, people talked slower than in the first part of the LATE strategy (almost significant, p =0.052). However,

Strategy	Mean	SD	Significance
Early half 1	3.05	0.89	0.011*
Early half 2	3.33	0.95	
Direct half 1	3.18	0.86	0.169
Direct half 2	3.28	0.90	
Late half 1	3.35	0.64	0.202
Late half 2	3.18	0.75	

Table 5.4: Comparison of speech rates within a strategy. * = p <0.05

Strategy 1	Mean	Strategy 2	Significance
Early half 1	$<$	Direct half 1	0.578
Direct half 1	$<$	Late half 1	0.200
Early half 1	$<$	Late half 1	0.052
Early half 2	$>$	Direct half 2	0.413
Direct half 2	$>$	Late half 2	0.154
Early half 2	$>$	Late half 2	0.045*

Table 5.5: Speech rate comparisons across strategies. * = p <0.05

in the second part of the EARLY strategy people talked faster than in the second part of the LATE strategy (p <0.05). It is no surprise that the speech rates in the first halves of the different strategies are not significantly different from each other since speakers need some time to adapt to the interlocutor's behaviour. So, when people were adapted to a certain strategy, they talked faster when they were interrupted than when the agent left silences between the turns. Table 5.4 and 5.5 show that users increased their speech rate when the EARLY strategy was applied (Table 5.4, $t(18 = -2.84, p = 0.11)$ and after doing so, they spoke significantly faster than when the LATE strategy was applied (Table 5.5, $t(18) = 2.16, p = 0.045$). Interestingly, Table 5.4 also shows that in the first half of the conversations, people spoke slower when the interviewer used the EARLY strategy than when it used the LATE strategy. A possible cause for this is that the interviewees were surprised when the interviewer started early and did not know how to handle the situation until the second half of the conversation.

Turn durations We also observed that when people were often interrupted, they not only spoke faster, they also had shorter speech turns. To verify this observation, we extracted all user turns that were not interrupted by the agent, and assumed that those turns were finished. After measuring the duration of these turns we compared them to each other. The results are shown in Figure 5.5.

This figure shows that the durations of the user turns in the different scenarios are

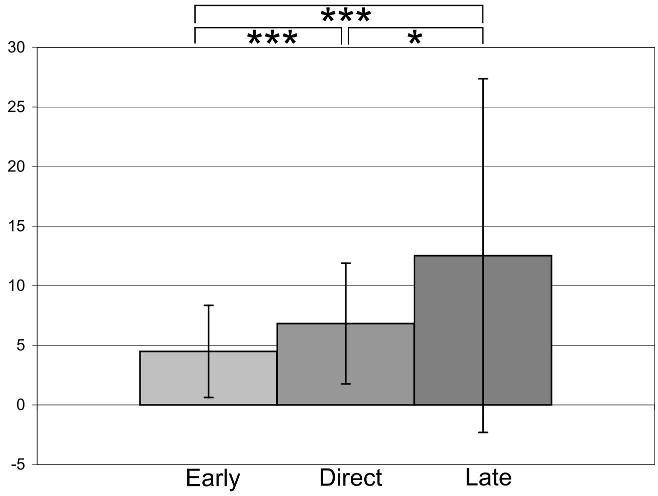

Figure 5.5: Non-interrupted turn durations. * = p <0.05, ** = p <0.01, *** = p <0.005

all significantly different. When interrupted often (in the early strategy) people made much shorter turns, and when the agent was very slow in taking the turn the users made much longer turns.

5.3.4 Agent gender

We observed that the voice that was used – male or female – made a big difference in the perception of the user. In the analysis of start-up strategy this effect was filtered out by using an equal number of male and female voices. However, the differences between the voices are still interesting, which is why we analysed the differences between these voices as well. Figure 5.6 shows the results.

This figure shows that the male voice was rated higher in Factor 1 (agreeableness), lower in Factor 2 (assertiveness) and lower in the arousal scale. This means the male voice was perceived, among other things, as more friendly, positive, polite, submissive, shy, and less aroused. The female voice was perceived, among other things, as more cold, aggressive, negative, dominant, bold, and aroused. This may appear strange, but the results probably say more about the voices used than about gender in general.

5.3.5 Further analysis

We also did some further comparisons involving the gender of both participants and the turn-taking strategy, but we only found some minor differences. More data is needed to verify the results, but some trends are already visible. We will mention the results here for the interested reader, but will not elaborate on them any more.

We also performed some other gender comparisons in the data which showed only minor differences. Robinson and Reis (1989) explain that an important factor could be the difference between the genders of the conversation participants – same-sex or opposite-sex. To study this we compared the gender of the user with the gender of the agent. However, only two minor results were found. Male users rated male

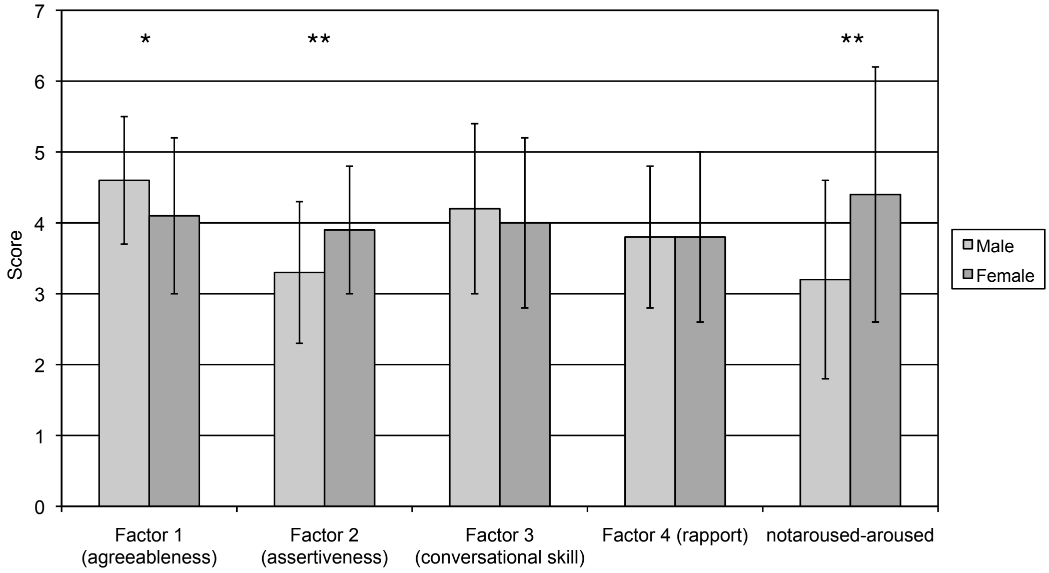

Figure 5.6: The results of the different genders. * = p <0.05, ** = p <0.01, *** = p <0.005

agents significantly lower ($p < 0.05$) in the notaroused-aroused scale than they rated female agents. Also, female agents were rated significantly lower ($p < 0.05$) in Factor 4 (rapport) by male users than they were rated by female users.

We were also interested in the combination of the start-up strategy and the gender of the user. For example, to check whether a male user perceived an agent using the EARLY strategy differently than a female user. However, no significant results were found. Another interesting combination is the start-up strategy and the gender of the agent. A male agent using the EARLY strategy might be perceived differently than a female agent using the same strategy. Only one result came out of this. A male agent using the DIRECT strategy was perceived significantly higher in Factor 1 (agreeableness) than a female agent using the same strategy.

The final thing we checked was the combination of user gender, agent gender and start-up strategy, but we only found one significant result here. A male user rated a male agent using the LATE strategy significantly lower ($p < 0.05$) in the notaroused-aroused scale than a female agent using the same strategy.

5.4 Summary

In this chapter, we presented our second study about the perception of different turn-taking behaviour. However, in contrast to the study in the previous chapter, the participants were actively engaged in the conversation. They talked with a computer in the role of an interviewer asking questions to the participants. The turn-taking behaviour of the computer was controlled by a human wizard, who chose whether each next question started before the participants finished their answer, directly afterwards, or after a small pause; these are, respectively, the strategies EARLY, DIRECT and LATE.

Each participant talked three times with the system, varying (counterbalanced) the topic of the questions, the gender of the system's voice, and the turn-taking strategy used. Afterwards, the participants completed a questionnaire containing 27 semantic differential scales. A manipulation check was performed to verify whether the human wizard performed the three turn-taking strategies correctly. We verified that in the resulting conversations, the EARLY strategy resulted in the lowest mean duration of silence between the turns and the highest number of overlaps, while the LATE strategy resulted in the highest mean duration of silence and the lowest number of overlaps.

Analysing the final results, we found that the three turn-taking strategies were indeed perceived differently. The interviewing agent using the EARLY strategy was perceived as most assertive and rude and least agreeable, and having the least conversational skill and rapport. Also, the agent using the LATE strategy was perceived as least assertive and most agreeable and respectful, and having the most conversational skill and rapport. We also found that different turn-taking strategies affected the speech behaviour of the user. Participants talking with an agent using the EARLY strategy increased their speech rates during the conversation and produced the shortest (finished) turns.

In the next chapter, we will present the conclusions and discussion of our two studies.

6 Conclusion and Reflection

In the previous chapters we explained that turn-taking could be used in different ways, and that each way has a certain effect on the perception that the other person has. In this chapter, we summarize and discuss the studies and their results. We also provide some thoughts on how the results can be used, and how they compare with the literature. Finally, we will discuss possible directions for future work.

6.1 Turn-taking

In Chapter 3, we tried to make clear that there is no 'best' strategy for turn-taking that can be used all the time. Researchers who develop dialogue agents often make them start as soon as possible after the end of the previous user turn, with as little overlapping speech as possible. But in human-human conversations, a non-trivial number of turn changes do not follow this pattern. Especially if the two interlocutors know each other, a lot of turns start while the other person is still talking, and silence gaps between turns are not always extremely short. This indicates that starting too early or too late happens in human conversations too, and is not always bad per se. Of course the default turn-taking behaviour of the virtual agent should be to start talking when the user is not speaking anymore without letting him or her wait too long for a reply, but deviations in this behaviour are a natural phenomenon and should not be avoided at all costs. They could even be used to the advantage of the agent.

We explained that differences in turn-taking behaviour can be very informative about the person using them and about the conversation itself. We showed that in certain types of conversations particular turn-taking behaviour occurs more or less frequently, and that different turn-taking behaviour is also perceived differently. In the context of the SEMAINE project — where we aimed to build four different virtual agents with different 'personalities' — we could use these differences in turn-taking behaviour to convey different impressions to the user. By using different turn-taking strategies we could influence how the agent was perceived by the user. But in order to use that, we needed to know how differences in turn-taking behaviour change the user's perception of the agent.

6.2 Passive Study

In Chapter 4 we described a basic conversation simulator that can generate artificial conversations that closely resemble human face-to-face conversations. The speech of the artificial participants is vocalized with unintelligible fragments of sentences from a real conversation, edited to filter out the content while keeping the prosody. We used this simulator to generate a number of conversations in which strategies for timing the beginning of a turn and strategies to deal with overlapping speech were varied. We showed, through an 'agent perception study', how these variations in turn-management changed the impressions that people gained of the agent as they listened to the various conversations. The study shows that the manipulation of turn-taking strategies can lead to different perceptions of an agent on personality scales, interpersonal scales and emotional scales.

The turn-taking strategies that were varied were the start time of a turn and how to deal with overlapping speech, that is, when both agents talk at the same time. The start time of a turn, in relation to the end of the other agent's turn, can be EARLY, DIRECT or LATE, which consecutively means that the next turn starts before the end of the other agent's turn, directly when it finishes, or after a small pause. The strategy to deal with overlapping speech can be either to STOP speaking, to CONTINUE normally, or to continue with a RAISED voice. When varying these turn-taking strategies, we gave each agent one start-up strategy and one overlap strategy.

We found that the EARLY+RAISED strategy received most negative ratings: it was perceived as most negative, unfriendly, rude, distant, cold and aroused. On the other end of the spectrum, the DIRECT+STOP strategy received most positive ratings: it was perceived as most positive, friendly, respectful, pleasant and warm. The most polite strategy (LATE+STOP) was perceived as most passive and submissive.

When looking at the individual strategies, we found that starting EARLY was seen as significantly more unfriendly, disagreeable, rude, cold, and active. Starting LATE was perceived as significantly more submissive and passive. For the overlap resolution strategy, STOPPING when overlap was detected was perceived as significantly more warm, passive, and submissive, and continuing with a RAISED voice was seen as significantly more negative, aroused, unfriendly, disagreeable, rude and unpredictable.

When looking at which start-up strategy was most pleasant, positive or friendly, unfortunately no strategy had significantly better ratings than the others. However, the DIRECT strategy does score highest for all three dimensions. Even though this result was not significant, it does indicate that in this setting the DIRECT strategy was preferred.

Most results are not really unexpected. It is a kind of social rule that you should let someone finish his or her turn before you start speaking, which means that it is very likely that an interruption is perceived as rude and unfriendly. And trying to keep the turn by speaking louder is also not very social. However, we should point out the shortcomings of this method. In our recordings, the raising of the voice was very apparent, but chances are that in a real conversation raising your voice is done in a more subtle way. Little research has been done on exactly *how* people behave when raising their voice when they want to keep the turn. Also, for interruptions, the fact that content was stripped from the recordings is also a downside. By not providing the context of the conversation and details of the content of the interruption, the rater was forced to treat the speech that start too early as an interruption. Factors that can make an interruption acceptable — for example, answering a question that is almost finished, an enthusiastic response, a warning for an angry dog, or finishing the sentence for the current speaker followed by a takeover of the turn — are no longer present in the recordings.

But if unintelligible speech strips away elements that can make an interruption acceptable, why make the speech unintelligible? One could argue that removing the content of the speech also removes an important source of information people use to formulate their impression. But this is actually precisely why we removed it. By taking away the actual content, we forced the raters to focus on the timing and prosody of the speech. As a result, our results describe how users perceive the different turn-taking strategies in a context-less setting. Of course, context can change the perception, and the same turn-taking behaviour can be perceived completely differently in different contexts, but this study functions as a starting point with as little context as possible. Additionally, another important, although not the main, reason is that it makes the conversation simulator and the study much simpler. When the content is not stripped away, the simulator has to take care that the generated conversations 'make sense', which is potentially a lot becomes complex. Also, with the content intact, the study itself gets more complex; the topic and the sentences affect the user's perception, which means that it has to be varied and counterbalanced. For all these reasons, we argue that using unintelligible speech was a viable choice which still produces useful and valid results.

Another tricky element in this study was the behaviour of the system-agent. This is the agent that all scripted agents talked with, and that used random turn-taking behaviour each turn. One can argue that the randomness of the system-agent affected the perception of the scripted-agent, because the behaviour of the scripted-agent depends on the turn-taking behaviour of the system-agent. However, we argue that, since the system-agent used random behaviour each turn, each scripted-agent encountered all possible turn-taking strategies during a conversation. This is the most fair solution and gave all scripted-agents similar conditions.

It should be noted though that unfortunately only 10 participants joined the study. This makes the results a bit less reliable, which is also why the ACTIVE experiment was conducted: to improve the reliability of the results.

6.3 Active Study

In Chapter 5, instead of a study in which the human rater was a bystander of the conversation, we let the rater actively participate in the conversation. With a Wizard-of-Oz setup, we simulated a conversational interviewing agent, and the start-time of the next question of the agent, which was controlled by the wizard, was determined by the turn-taking strategies we were testing: EARLY, DIRECT or LATE. In the study, we varied the start-up turn-taking strategy, the topic of the questions, and the gender of the system's voice. Although it is not easy to guarantee that a Wizard uses a certain strategy consistently, the analysis of the recordings revealed that the three different strategies were applied accordingly in this experiment. After each conversation, the participants had to complete a questionnaire about how they perceived the system, containing 27 semantic differential scales that captured not only interpersonal and affective dimensions, but also dimensions that were more social-skills related and concerned interviewing capabilities.

Based on the results we found, we can conclude that an agent that used a certain turn-taking strategy could indeed influence the impression that a user had of this agent. Starting too early (that is, interrupting the user) was mostly associated with negative and strong personality attributes: agents were perceived as less agreeable and more assertive. Leaving pauses between turns had contrary associations: it was perceived as more agreeable, less assertive, and created the feeling of having more rapport. The agent's voice played a role in the results too. In general, we can say that the male voice was perceived as more agreeable and less aroused than the female voice. However, this effect could be more related to the quality of the synthesized voices than to the gender of the agent. Since we only used two different voices for each gender (one Dutch and one English) it is very hard to generalize these results to gender. Previous studies also report on relations between gender and interruptions and interpersonal perceptions of interlocutors — for example, females who interrupt would be penalized more than male interrupters — but we did not find such effects in our data. This was mainly because our prime interest was the turn-taking strategy.

Finally, an analysis of the user's turns in interaction with the interviewing agent showed that turn management not only influences the impression one has of an agent, but it also influences the response behaviour of the user. The user seems to 'adapt' to the interviewing agent's turn-taking strategy. For example, users spoke significantly faster when the EARLY strategy was applied than when the LATE strategy was applied. Furthermore, the results showed significant differences in turn lengths between all strategies: the EARLY strategy was associated with short turn lengths while the LATE strategy was associated with long turn lengths.

We have to keep in mind though that the results were obtained in the interviewing domain, and that some findings might not generalize to a 'free-talk' conversation in which dialogue partners can talk about anything they like, or in a setting in which both dialogue partners can ask questions to each other.

This leads to an important question, namely how the interview setting influences the results, and how specific these results are for this domain. As stated before (see Chapter 5), interviews are usually governed by very strict rules. The most important aspect of the interview is the story of the interviewee. The interviewer wants as much of this story as possible, and the interviewee wants to tell as much as he or she can. This creates a certain imbalance in the conversation, where the interviewee is more important than the interviewer. Because of this, the interviewee's behaviour — for example interrupting the interviewer or ignoring attempts to take the turn — is more easily accepted than the interviewer's behaviour. And vice versa, the interviewer's behaviour is much more strict, and unwanted behaviour is far less acceptable. In short, a rude interviewee does not break down the conversation that easily (the interviewer still wants the story), while a rude interviewer has a much higher chance of breaking down, probably with the interviewee walking away.

So, the interviewer is socially required to behave very politely, and impolite behaviour is punished much harsher. The results of the ACTIVE experiment show this as well: interrupting the interviewee was perceived as more rude, less agreeable and shows less conversational skill than giving the interviewee more time to finish by starting late. This means that this setting is an extreme case, and we argue that the results of this study can at least be generalized to other polite conversations or settings, such as a virtual receptionist or guide. We have to mention though that we do not have any evidence for this claim, and leave this question open for further research.

Another debatable aspect of the study is the influence of the wizard. Since we did not use a blind wizard, he knew the aims of the study, and could therefore potentially influence the outcomes of the study. Also, since we only used one wizard, the outcomes of this study are influenced by the wizard's interpretation of the intended behaviour of the turn-taking strategies. This could be a potential problem, but in our study the only aspect of the conversation that the wizard could control was the start-time of the next question; there was nothing else he could influence. And with the manipulation check (see section 5.2), we made sure that the performed behaviour of the wizard was as intended.

6.4 Discussion

We have performed two studies to try to find out how different turn-taking strategies change the user's perception. But studying changes in perception is hard to do well, with the result that both studies have their shortcomings. The PASSIVE study was conducted with only 10 participants, the speech was incomprehensible, and the participants were bystanders in the conversation. The ACTIVE study was conducted in the interview domain, arguably limiting the generalizability to this domain only. However, both studies were looking for the same effects, but with different approaches. Limitations in one study do not exist in the other, and each study has a unique approach. As a result, combining the two studies together strengthens the validity of our findings.

When comparing the results of the PASSIVE and the ACTIVE study, we can see that the impression one has of an agent does not radically change when a situational context (i.e., topics to talk about in an interview setting) is provided and when one is an active conversationalist in the experiment, as opposed to being a passive bystander listening to unintelligible conversations. In general, the results from the PASSIVE study can be transferred to the ACTIVE study. A noticeable difference is that the ratings in general in the ACTIVE experiments are more extreme; users rated the ACTIVE system lower in the EARLY strategy and higher in the LATE strategy. Here, lower ratings correspond with a lower perceived level of assertiveness, less conversational skill, and less rapport. A possible explanation for these more extreme ratings is that the subjects who were in interaction with the interviewing agent were more engaged. They punished and rewarded the agent more because they *underwent* the effects of the strategies themselves, rather than *observing* these.

Another notable difference between the results of both studies is that in the ACTIVE study the lowest rating and the highest rating were always given to the EARLY or the LATE strategy. However, in the PASSIVE study, some ratings have the highest or the lowest value with the DIRECT strategy. Unfortunately, these values are not significantly higher or lower, but they do indicate subtle differences. In the PASSIVE study, the DIRECT strategy was perceived as more positive, friendly, pleasant, attentive, warm, and responsible than the other two strategies. In contrast, in the active study, the highest values of the corresponding factors — agreeableness and conversational skill — are perceived with the LATE strategy.

The same can be seen when comparing the results of the Agreeableness-factor (see Figure 4.4 and Figure 5.4). The factors of the PASSIVE and the ACTIVE study are grouping similar semantic scales, and four of these scales appear in both factors: disagreeable–agreeable, unfriendly–friendly, negative–positive and cold–warm. When comparing the results, in the ACTIVE study the lowest value was perceived with the EARLY strategy, and the highest value with the LATE strategy. However, in the PASSIVE study, the highest value was perceived with the DIRECT strategy, while the LATE strategy had a slightly lower value (though not significantly). This is probably explained by the difference in context. Since the interviewing setting was very polite, users were less lenient and wanted more time to finish their turn. This probably means that the 'best' time to start speaking (when it is perceived as most positive, friendly, pleasant, and so forth) shifts to a later time. Although very interesting, it was out of the scope of this study. The other factors cannot be compared, since the semantic scales that they consist of are not similar enough.

An important aspect of our two studies is that they looked at global perception: the holistic impression that participants have at the end of the conversation. However, different turn-taking strategies can also have more local perception changes, for example a person who looks very enthusiastic because he starts his turn early. In contrast with global perception, local perception changes are much harder to measure and are affected by many more factors, for example the context of the conversation, the environment, the content of the speech, and other signals that the interlocutors use while performing turn-taking behaviour.

6.5 Applying the results

We started these studies because we wanted to know how we could use turn-taking as a tool to change the perception that people have of a virtual agent. We have shown that different turn-taking strategies can indeed change this perception, and can even change the behaviour of the user. But an obvious follow-up question is: how can we use this knowledge?

For Sal, the differences between the four characters are very clear. Spike, the aggressive character, is supposed to be active and negative. He would therefore start speaking early, and will probably raise his voice when overlapping speech is detected. This behaviour will make the user perceive Spike as more rude, assertive, cold, unfriendly, dominant, and active, and less agreeable: exactly what we want. Obadiah, the depressed, negative and passive character, is on the other end of the spectrum.

His turn-taking behaviour will be very late, and when overlapping speech is detected he will simply stop speaking. Users will then perceive him as more submissive and passive. Unfortunately, it could also make him seem more agreeable and respectful, which is not what we want. Perhaps if we increase the pause duration even further the positive traits will disappear, but this is something for future research. Poppy — the happy, active and positive character — and Prudence — the pragmatic and positive character — will use a turn-taking strategy that is somewhere between Spike's and Obadiah's. By starting directly, users will get a friendly and warm perception of the agent, without noticing the negative traits that appear when using a more extreme strategy (starting early or late). Differences between Poppy and Prudence are probably very small.

Applying the results on these four characters is pretty straightforward, but that is mainly because their 'personalities' are very clear-cut. But a lot of virtual agents are not angry or depressed by default, so how can we use the results for virtual agents that have a less obvious personality?

In a formal setting — for example, where the virtual agent performs some kind of social service to the public, such as a virtual guide or receptionist — people expect the virtual agent to behave in a very polite way. This means that interruptions are not acceptable, and it would be better to wait just a bit longer to avoid interruption. This also improves the perception of conversational skill of the agent. There is of course an optimal time to wait, and waiting too long after the user's end of turn will only make the agent seem unresponsive.

In a more general setting, the results of these studies can be useful too. For example, when the agent notices that the user is getting too dominant in the conversation, it could make itself more dominant to restore the balance, for example by interrupting the user or raising its voice when it is interrupted by the user. Or perhaps the agent wants to increase the user's perception of how attentive it is, for example when the user tells something interesting. By reacting more directly to the user the agent increases the perception of attentiveness.

Instead of using the different turn-taking strategies to change the user's perception of the agent, it can be used to strengthen the effect of another behaviour. For example, when disagreeing with the user, this can be augmented with starting to speak early, amplifying the disagreement. And when the agent wants to respond enthusiastically to the user, responding just before the user's turn is finished boosts the effect by increasing the perception of arousal.

This shows that there are numerous examples of how the results of these studies can be used as a useful tool to play with the perception that the user has of the virtual agent.

6.6 Comparing the results with the literature

In Chapter 3 we showed summarized several studies from the literature that look at the perception of turn-taking, especially interruptions. But how do their results compare to ours?

Trimboli and Walker (1984) studied the difference between cooperative and competitive conversations, and found that competitive conversations contained significantly more interruptions. In competitive conversations, it is likely that the participants do not agree with each other, but are constantly trying to get their point across. This means that the participants have a low agreeableness and a high assertiveness. Our results show that agents that use the EARLY strategy, which leads to more interruptions, are also perceived as having a lower agreeableness and a higher assertiveness.

Goldberg (1990) explained that interruptions were originally linked to displays of power and control, but she argues that they are also linked to a higher rapport, cooperation, and enthusiasm. Our results partially show this too: agents using the EARLY strategy were perceived as having a higher assertiveness — which contributes to the displays of power and control — and also as being more active and aroused — which contributes to the enthusiasm aspect. However, in our study we can only confirm that the EARLY strategy was perceived as showing low rapport.

Robinson and Reis (1989) show that interrupters in a conversation are perceived as more assertive, more masculine and less sociable. Our results also show that the EARLY strategy was perceived as more assertive and with having less conversational skill. Interestingly, when looking at the gender, we perceive something different. We did not directly study perceived masculinity, but our second study did contain male and female voices. In this study, the EARLY strategy was perceived as less agreeable and more assertive. However, when looking at the differences between the voices, the male voice was perceived as significantly *more* agreeable and *less* assertive, which is contrary to the EARLY strategy. However, this is not a direct comparison between perceived masculinity and turn-taking strategy, and it is also a possibility that these results are more related to the specific voices instead of the gender.

6.7 Future work

We can make several recommendations for future work.

- It would be interesting to look at more subtle differences between turn-taking styles and to apply these as strategies. In our experiments, each strategy was uniformly applied to all turn starts, that is, in the EARLY strategy all turns start too early, and in the LATE strategy all turns start too late. It would be interesting to identify more turn-taking styles, for example by mixing several strategies. One could think of a strategy that alternately starts directly or late, or one that interrupts only once in a conversation.

- As already mentioned in section 6.4 we only looked at the global perception of the user. It would be very interesting to look at more local perception changes. For example, what changes in the user's perception when the agent interrupts the user one time, or when the agent waits for a long time once?

- The role of context and type of conversation in turn management deserves more attention. Although we have identified some results that are transferable between the PASSIVE and the ACTIVE experiment, it still remains an open issue how large the influence of context and type of conversation — for example a work discussion, a conversation between friends, or someone asking for directions — is in these types of perception studies. Does the perception change in different contexts? Do people tolerate more or less? And does the timing of the highest and the lowest rated strategy change? For example, is the best pause duration between turns different in formal or informal conversations?

- Finally, the turn-taking strategies should be implemented through a state-of-the-art real-time end-of-turn detector and evaluate the 'real' effects of the strategies on perception through interactive experiments. One of the challenges in this will lie in developing an real-time end-of-turn detector that can adopt the EARLY and DIRECT strategies.

Part III

Response selection

7

Response selection in Sal

This part of this thesis is about the methods we use to select appropriate responses in Sal. In this chapter, we will first introduce common techniques that are often used for response selection, and we will explain when these methods are usually used. After that, we will describe the response requirements of Sal, how this affects the response selection and which techniques we can use. Finally, we will explain our first iteration of creating selection rules, in which we hand-crafted the rules ourselves.

In the next chapter, we will explain a more data-driven approach, in which we use Wizard-of-Oz data to make the agent learn the response selection rules.

7.1 Response selection methods

In this section we will describe some common techniques for response selection that are often used in virtual agents or other dialogue systems.

7.1.1 Finite state machines

The simplest method to model when to say what is the finite state model, and is described in more detail by Bui (2006). This model contains every possible state of the conversation, the response that the agent gives in that state, and the transitions between all states. An example of such a model is shown in Figure 7.1. A train information agent using this model starts in state 1 and asks “Can I help you?”. If the user confirms this, but does not specify any details about the trip, it will go to state 2 and ask the departure place of the user. If the user initially responded with “Yes, I want to travel from Utrecht to Groningen”, then the system changes its state to state 4 and continues by asking when the user wants to travel.

However, there are some downsides to this approach. Every possible state in the conversation should be predetermined, and the responses of the agent as well. This makes the system inflexible: it takes time to extend the model, and changing the conversational structure means that numerous, if not all, models should be redesigned.

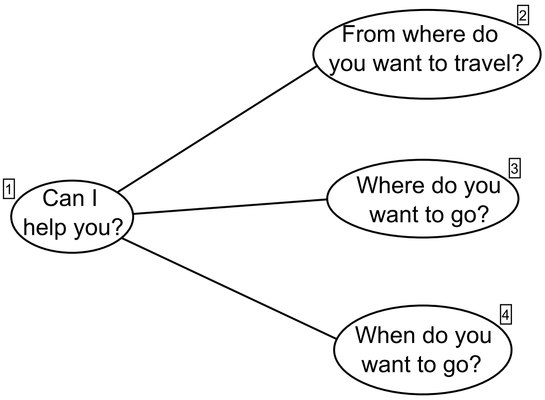

Figure 7.1: Example fragment of a finite state model of a conversation.

7.1.2 Frame based

A more flexible method is the frame-based approach (see Bui (2006)). The agent uses *forms* or *slots* to keep track of the information it needs in order to perform its task, and tries to fill these slots by asking the user for the information. For example, for the train ticket agent, the slots could look like this:

```
[Travel:
  Origin = Utrecht
  Destination = .
  Date = today
  Time = .
]
```

When it deduces information from the user's response the agent fills in a slot (for example the destination). Using the current state it decides what to ask the user next, for example at what time the user wants to travel, or perhaps a confirmation if the agent is not sure ("You said you want to go to Groningen?").

An important aspect of response selection with frame-based models is understanding the user's input. With a 100% correct understanding, the only thing the agent has to do is ask the user for the information it could not deduce itself. Unfortunately, perfectly understanding the user is difficult, which means that agents typically have to verify their understanding, either explicitly ("Did you mean 'Groningen'?") or implicitly ("And when do you want to travel to Groningen?"). Too many verifications disrupt the flow of the conversation and make it awkward to talk to the agent.

7.1.3 Information state based

A method that makes it easier to use contextual and multi-modal information is the information state based approach (Larsson and Traum, 2000). This approach tries to overcome the limitations of the finite-state and frame-based models. According to Traum and Larsson (2003), the information state approach consists of five main components: the components in the Information State (IS) — beliefs, intentions, conversational structure, etc — and their formal representation, rules that update the data in the information state, dialogue moves that trigger these update rules, and an update strategy that decides which rule to apply when.

The finite state approach provides a framework to store contextual information in a structured way and create update rules that modify this information. The original information state approach was extended by Traum and Rickel (2002). In their approach, the authors use a multi-layer dialogue model with an information state in each layer to store its current status.

However, specifying the actual rules (how to respond on the content of the IS) is more time-consuming than in the finite-state and frame-based models. A popular solution is to use corpora to learn how to respond to changes in the information states, for example by using Markov Decision Processes (Levin et al., 2000), Q-learning (Scheffler and Young, 2002), or with Bayesian Networks (Wai et al., 2001). But even when learning dialogue strategies from data, when the domain gets too extensive the amount of required work quickly increases. The system will always need domain-knowledge to understand the user before it can perform its tasks.

7.1.4 Information retrieval approach

This approach views the problem of response selection as an information retrieval problem, in which the user's utterance is the query, and the appropriate response is the most fitting result. The approach is often used in question-answer systems. For simple fact-based questions, it is sufficient to convert the user's question to a database query, run that query on a database that contains all the facts and create a response with the results. But this approach can also be used for non fact-based questions, as demonstrated by McCauley and D'Mello (2006). They store all possible answers of their agent MIKI in the database, and for each answer a list of possible questions. Using Latent Semantic Analysis (Landauer and Dumais, 1997) they find the stored question that is most similar to the user's question, and they return the corresponding answer.

7.2 When to use which response selection method

In the previous section we described some common approaches of response selection. However, when to use which approach depends mostly on the context of the system

If the domain contains a lot of small and structured sub-dialogues, then finite state machines (FSMs) are a good choice. FSMs are easy to create, but the time this requires increases quickly when the conversation gets more complex. Also, it requires the developers to flesh out all possible conversations beforehand. This means that finite state machines are mostly suitable for small and very structured domains. For example, MACK (Cassell et al., 2002) is a virtual receptionist that can answer questions about the departments, its projects and people, and it can give directions. Each task is straightforward and small, and the number of tasks is also not that great. The same holds for Marve (Babu et al., 2005), who can also take and deliver messages to people it knows. These tasks always follow the same structured dialogue, and finite state machines are therefore the ideal approach to model these sub-dialogues.

Task oriented dialogue systems are, as the name implies, systems that have to perform some kind of task for the user, such as selling a train ticket or giving restaurant suggestions. Conversations in these domains are usually characterized by an information gathering phase, in which the system gathers all necessary information from the user, and a result phase, in which it performs its task. An obvious approach would be to use the frame-based method, in which the necessary information is stored inside slots. As long as there are empty slots, the agent can ask this information of the user. When all slots are filled in, it can use the gathered data to perform its task. A good example of a frame-based task oriented dialogue system is Karen (Nijholt and Hulstijn, 2000), a virtual receptionist of a music center. Karen can provide information about performances and sells tickets to visitors. Another example is the public transport information system described by Van Zanten (1996), which provides timetable information. If the task is simple enough, a finite state machine is feasible too.

Question-answer systems are meant to answer the user's questions. Often these systems are only required to give an appropriate answer and not to maintain a complete coherent conversation. It is possible to use a finite state machine to model the question-answer pairs, but there are some reasons against this. First of all, it requires one new state for each question-answer pair, so the more questions a system can recognize and possible answers it can give, the more states (and work) are needed. Secondly, the biggest problem in question-answering is understanding which question is being asked. However, Cassell et al. (2002) show with MACK that if the domain, and thereby the number of questions and answers, is small enough, then finite state machines can work.

Information state based methods work well in more complex situations, since all information about the user's question can be put in one information state. The system then needs rules that specify, based on the data in the information state, which answer to give. An example of such a system is IMIX (Op den Akker et al., 2005), a dialogue system that answers questions in the medical domain. Another approach is more database-driven, for example as used in Sergeant Blackwell (Leuski et al., 2006). This agent uses statistical techniques to find the best response, based on a database with possible question-answer pairs. An even more clear example of how information retrieval is used in question-answer systems is the system YourQA (Quarteroni and Manandhar, 2009). This system uses the internet to answer the user's questions, and uses information retrieval techniques to assess the quality of the found documents. The advantage over a database-driven method is that the developer only has to add more questions and answers to the database when extending the system, while the information state approach needs more hand-crafted rules. However, with the database-driven system it is more work to get more coherent conversations, since this method is designed to focus on coherent question-answer pairs.

When the type of system is not as straightforward as a question-answer system or a task-oriented system, then there are multiple solutions. The information state approach works well with multi-modal data and in complex situations, since one of its main strengths is combining data from different sources. Kopp et al. (2005) show with their museum agent Max that this approach is feasible in more complex situations, even though a lot of work was needed to create all the response selection rules. Therefore, a second solution might work better, namely using a mixture of approaches. A great example of this is the virtual agent H.C. Andersen, as described by Bernsen et al. (2004) and Bernsen and Dybkjær (2004). It uses templates when telling fairy tales, and finite state machines when going into structured, in-depth subdialogues. Using a mixture of approaches can combine the advantages of different methods while minimizing their shortcomings.

7.3 Which method for Sal

In order to determine which response selection approach is most suitable for Sal, we need the requirements of Sal. Remember from Chapter 1 that Sal consists of four different characters with different emotional states, and that Sal's goal is to motivate the user to keep talking for as long as possible, and also try to get him in the same emotional state as the current character. We decided on the following requirements for Sal:

- **No content** – The first requirement of Sal is that it uses none or as little as possible of the content of the user's speech. Instead of this content, Sal uses other information sources, such as prosodic information, facial expressions, head movements, and so on. This also means that, since the content of the user's speech is not taken into account, the domain in which the agent operates is completely open. Sal responds to how the user feels about what was said, independent of the actual topic. As a simple example, imagine that the user tells something while being very sad about it (they show depressed facial expressions and talk in a sad tone of voice). Without knowing what the user is talking about, Sal could know that he is not happy about it and could try to comfort him. This also means that Sal cannot answer questions, since then the content of the question should be known.

- **Being a listener** – It has to take the role of the listener. This means that it only occasionally says something, but keeps the other person speaking for as long as possible. When it does respond, it has to motivate the user to continue speaking, to react to what the user just said, or to introduce a new topic.

- **Affect user's emotional state** – When responding to the user, Sal tries to get the user's emotional state in the same state as the current character. For example, this can be achieved with its response (e.g. Obadiah saying "But just think of everything that can go wrong.") and its backchannel behaviour (e.g. Poppy giving a smile when the user tells that his vacation was amazing).

- **Limited number of responses** – Sal can only select a response from a finite list of responses, with about 140 responses for each character. In the HUMAINE project it was shown with a Wizard-of-Oz experiment, in which users had to talk to a wizard-controlled 'system' (Douglas-Cowie et al., 2008), that this list of responses was sufficient to sustain a conversation for some time, sometimes even in the order of half an hour. Therefore, we decided that Sal had to use the same responses. Some example responses for the four different characters (see Chapter 2) are shown in table 7.1.

Character	Response
Poppy	•It'll all be better tomorrow. •It's great to hear someone sound happy. •Well done! •What other good news do you have? •Give me just one happy thought and you'll feel better.
Prudence	•I'm sure that you can solve the problem, whatever it is. •It can't be that bad, don't you think you're over-reacting? •Don't get carried away! •Why should I think that's so good? •That sounds like a sensible attitude.
Obadiah	•There's not much you can do about it. •Tell me all the awful details again. •Well I suppose you have to make the best of it while you can. •Tell me about the last time you were really hurt. •It wears you down, doesn't it?
Spike	•It's a bloody disgrace. •I don't see why you're in such a good mood. •I suppose you really think you're something. •You've got to show people who's boss! •When was the last time you got angry at something?

Table 7.1: Example responses of the four Sal characters.

Sal is clearly not a task-oriented system; it does not perform a certain task for the user and it does not require particular information from the user. For this reason, a frame-based approach for response selection would not be useful for Sal. Finite state machines work best in structured and small domains, which is not really the case for Sal's conversations. However, there are some small phases during the interaction with Sal that are more structured, namely the start of the conversation and when changing characters. These phases always take the same 'routes' and would be perfect candidates to be modeled with finite state machines.

The information retrieval approach is more tricky. Sal is not a question-answer system in the most strict definition, but it does work with the same principles. Based on what the user has just said, Sal has to find the most appropriate response. Also, since Sal does not take content into account, it is very hard to achieve coherency that spans multiple responses. This means that, just as in a lot of question-answer systems, Sal only looks at the previous turn of the user when determining what to say, and the global context is less important. However, with this method it can be problematic to fill the database with the required information. All possible responses are known beforehand, but since Sal operates in an open domain, it would be impossible to list all possible ‘questions’ that lead to these responses. It would be possible to generalize the user’s behaviour to more abstract features and store these.

However, we decided to use the information state approach. This approach is known for its flexibility and it works very well with data from multiple sources, such as different modalities. The problem then remains to create good rules that specify when to select which response.

In the next section we describe the models we handcrafted, which was our first attempt to model the agent’s responses. These simple models worked occasionally, but not as well as we hoped. Therefore, we decided to let the computer learn the models using training data. We explain this approach in the next chapter.

7.4 Handcrafted models

In this section we explain the first attempt to model the dialogues for Sal. We did this by handcrafting dialogue rules using Flipper, an Information State based dialogue system (Ter Maat and Heylen, 2011). With this system, rules are written down in *templates*, which specify the preconditions and the effects of the rules. The rules were based on different aspects of the conversation that Sal could respond to, for example the arousal of the user, a long period of silence, and non-verbal behaviour. We will now describe the different models.

7.4.1 Start-up and character change models

Two parts of a conversation with Sal are more structured: the start of the conversation, and the switching of characters. At the start of the conversation, Sal greets the user, asks how he or she is today, and asks the user to tell more about it. When Sal detects that the user wants to change the character, or when Sal determines it is time to change the character, it can ask the user whether he or she wants to change, and to which character. If the user does not care, Sal can make a suggestion, which can be accepted or rejected. Because of the structured nature of these mini dialogues, they are implemented as state machines.

7.4.2 Arousal models

Arousal is one emotional feature that can be fairly reliably detected from the user’s speech. Using the arousal values, we created rules for Sal to respond to high or low values of arousal. Of course, the different characters respond differently to the arousal values.

Some example rules:

```
Prudence
Arousal = Low  -> ‘You seem a bit flat.’
Arousal = High -> ‘Don’t get too excited!’
```

```
Obadiah
Arousal = low  -> ‘Don’t get too carried away!’
Arousal = high -> ‘You’re a bit of a cold fish, aren’t you?’
```

7.4.3 Silence models

It often happens that the user does not know how to deal with Sal, which causes the user to not start talking after Sal makes a response. There are several ways in which we deal with this.

First of all, there is a chance that the user does not have anything more to tell about the current (unknown) subject. If Sal did not recently introduce a new topic, it will introduce one after the silence. Another possibility is that the user does not know how to respond to a question from Sal. To improve this, some questions have elaborations that can be presented after a silence. Two examples:

```
Poppy
‘I think you should feel really happy today.’
(after a silence)
‘Tell me what makes you feel really happy.’

Spike
‘When was the last time you got angry at something?’
(after a silence)
‘Don’t you ever get angry at the world?’
```

If these two approaches cannot be applied — that is, there is no elaboration on the last response and Sal recently introduced a new subject — then Sal will try to motivate the user to continue speaking. Again, two examples:

```
Poppy -> ‘Come on, tell me more.’
Poppy -> ‘Well?’
Spike -> ‘‘Go on, tell me your news!’’
```

7.4.4 Laughter models

Another detected feature that can be easily used for response selection is laughs. If laughter is detected, either detected in the audio or in the facial expression of the user, characters can respond differently to this behaviour:

```
Poppy -> ‘It’s great to hear someone sound happy.’
Prudence -> ‘Why should I think that’s so good?’
Obadiah -> ‘If you’re laughing, people won’t take you seriously.’
Spike -> ‘Stop laughing!’
```

7.4.5 Linked response models

In order to get more structured conversations that take some more context into account than only the previous user utterance, models were created that link a response of an agent to a previous response. Based on the previous response of the agent and some features of the user's utterance, a more coherent response can be selected. This kind of linking is an easy way to create a more structured conversation, without specifying much about the content. Several features of the user's turn are used to decide on the response. Currently, the features that can be detected are the length of the turn (short, normal or long) and an agreement or disagreement (basically, if the user says yes or no). With these features, all agent utterances that are questions and that could be answered with yes or no have been extended with proper responses. Some examples can be found in Table 7.2.

Character	**Previous response**	**Features of user's utterance**	**Response**
Poppy	Did things get better?	short disagreement	Cheer up!
Prudence	Would you say you have everything under control?	short agreement	Good. You obviously have your head screwed on.
Spike	Don't you ever get angry at the world?	short disagreement	You've got to show people who's boss!

Table 7.2: Example linked responses.

7.4.6 Last resort model

If all else fails, and if no model triggers any rule, then this model has to supply a response. These are mostly generic responses that fit in almost any context, such as 'Tell me more' and 'Go on'.

7.5 Conclusion

In this chapter we described several common response selection techniques, and we explained that we use a combination of the information state approach and finite state machines. Then we described the different handcrafted rules that we used in a first version of Sal. In the next chapter we will present a more data-driven approach, in which the computer uses data to learn selection rules.

8 Data-driven response selection

In the previous chapter we explained how handcrafted rules can be used to model the response behaviour of Sal. However, these rules were based on our ideas of appropriate responses, but these ideas might be wrong or insufficient. In this chapter we will explain the next step, in which we use a data-driven approach to create response models for Sal. We explain how we use videos to collect data of appropriate responses, and use these to train classifiers that learn how to give appropriate responses.

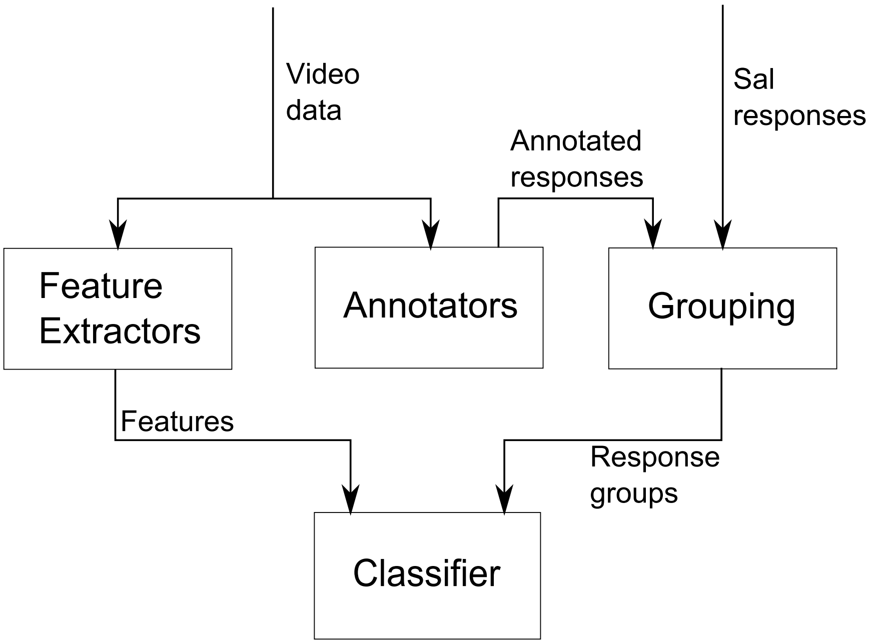

Figure 8.1: The structure of creating the classifiers. The video data is used to extract features and to get annotated responses. These suggested responses are grouped and outputted to the classifier.

Figure 8.1 shows the structure that we used in creating the classifiers. Using a corpus of human-human conversations in a similar setting as Sal (McKeown et al., 2010), we extracted features of the user's turn and appropriate responses that follow these turns, and used this as training data for the classifiers. However, in these conversations neither participants were restricted in what they said, while Sal is restricted to a finite list of responses. And since the classifiers need to produce responses from that list, the training data that was extracted from the corpus needs to contain the same responses. This was solved by having several annotators go through the data and select 3 appropriate Sal responses (response suggestions) to each turn in the data.

However, each Sal character has over 100 responses in its repertoire. Having the classifier learn to differentiate between them all requires a lot of data. Therefore, we clustered the responses in groups of responses that tried to say the same thing. We sent the extracted features from the users' turns and the groups of the suggested responses to different classifiers in different formats, because we did not know which classifier or format worked best.

In this chapter, we will first provide more details about the corpus we used and the features we extracted. We will then describe the annotation process and the different grouping approaches we tried. We will explain which classifiers we used and present the results. However, the performance of the classifiers is hard to measure, because responses that were not annotated could still be appropriate. Therefore, in section 8.6 we will explain the subjective evaluation in which we presented the results to human raters.

8.1 The corpus and its features

In this section, we will elaborate on the corpus and its features that we used in the classifiers. We used two types of features: the SEMAINE annotations that were manually added to the recordings, and automatically extracted features. The annotations would not be useful in a real system, because they are hard to acquire automatically by machines and therefore, a real state-of-the-art system would not be able to recover them. However, they could give insight into which annotations are useful to detect. For the automatically extracted features we only extracted audio features, because the video component that is used in Sal is only capable of processing live camera data.

8.1.1 The SEMAINE Corpus

In the data-driven approach, we used the SEMAINE corpus (McKeown et al., 2010). This corpus contains recordings of human to human conversations in the SAL-scenario, which means that one participant (the *operator*) takes the role of the listener. His or her goal is to keep the other participant (the *user*) talking for as long as possible and to try to draw the user to the emotional state of the current character that the operator is playing. These characters are the same as the ones described in Chapter 2, namely optimistic Poppy, pragmatic Prudence, depressed Obadiah, and aggressive Spike.

For the SEMAINE corpus, 20 conversations were recorded, in which a user talked with an operator for an average duration of 20 minutes. During these conversations, the operator played out all four characters. The participants were sitting in different rooms and saw each other through teleprompter screens. Four cameras were placed behind semi-reflective mirrors to record the user and the operator frontally: one color

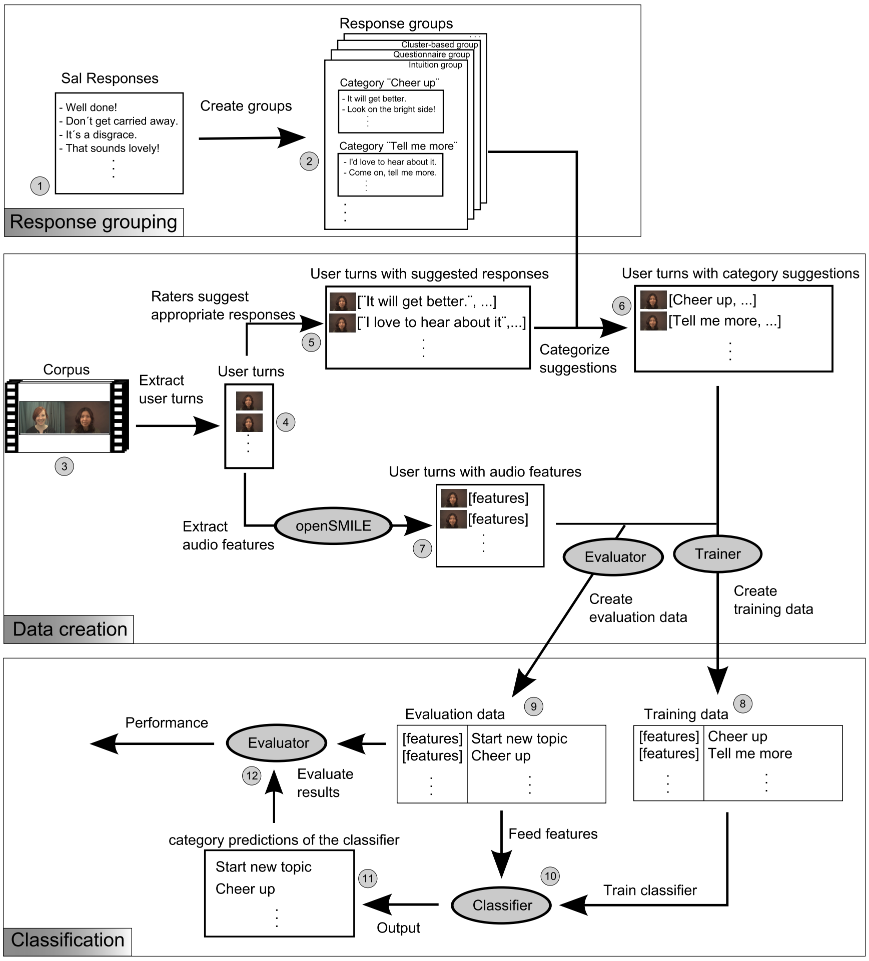

Figure 8.2: A detailed diagram of the data-driven response selection. 1. All predetermined responses that are available in the Sal system (section 7.3).
2. All responses of the Sal system are ordered using different groups (section 8.3). In each group, the responses are divided into categories.
3. The Semaine corpus that was used as training and evaluation data (section 8.1.1).
4. Fragments that contain user turns are extracted from the corpus.
5. Raters use these fragments to suggest up to three appropriate responses for each user turn (section 8.2).
6. Using the different groupings, the response suggestions are changed to the corresponding categories, which results in a list of suggested categories for each user turn.
7. Using openSMILE, audio features are extracted from the user turns (section 8.1.3).
8. Using different Trainers (section 8.4.1), the features and the suggested categories are converted to a training data set that the classifier can use.
9. Using different Evaluators (section 8.4.2), the features and the suggested categories are converted into an evaluation data set.
10. The training data is used to train classifiers (section 8.4.3), and then the evaluation data is fed to these classifiers.
11. The classifiers predict categories for the user turns in the evaluation data set.
12. Using the same Evaluators (section 8.4.2), the predicted categories are compared with the suggested categories to calculate the performance (section 8.5).

and one grey-scale camera for each participant. There was a grey-scale camera aimed at the side of the user too. As the cameras were placed behind the semi-reflective mirrors, the participants could look each other in the eyes during the conversations. Two microphones were placed in each room, one on the table in front of the user or operator, and one on the head of the participant. The recorded data was synchronized, and the videos were compressed.

The operators were instructed to play each character as naturally as possible, without being tied to a fixed script. Users were not allowed to ask questions. Changing the character could be requested by the user, for example when they got bored or annoyed. Also the operator could also request a character change when enough time had been spent with the current character.

8.1.2 SEMAINE Annotations

The SEMAINE data contains trace-style continuous annotations — in which the value of the annotation is continuously determined by 'tracing' the current value with a slider while watching the video — in 31 different dimensions related to emotions, epistemic states and dialogue intentions. Every recording is annotated on five core dimensions: Valence, Activation, Power, Anticipation/Expectation, and Overall Emotion Intensity. Fontaine et al. (2007) argue that the first four dimensions are best suited to capture affective colouring in general, although they name the dimensions evaluation-pleasantness, activation-arousal, potency-control, and unpredictability. Valence and activation form the basis of the different characters, where valence specifies whether the participant is positive or negative, and activation how aroused the participant is. We designed the four characters of Sal to fit these descriptions, so Poppy is designed to be positive-active and Obadiah to be negative-passive. Power is characterized by appraisals of control, leading to feelings of power or weakness, and anticipation/expectation is characterized by appraisals of novelty and unpredictability. The final core dimension is that of Overall Emotion Intensity, to measure the overall level of the emotional intensity of the participant.

Besides these five core dimensions, each annotator annotated an additional four (at least) out of 26 possible dimensions, depending on which dimensions had at least one obvious case in the recording. The 26 possible dimensions are divided into four categories, as shown in Table 8.1: basic emotions, epistemic states, interaction process analysis, and validity. The validity category tries to highlight problems in the conversations, for example a breakdown of the engagement or when the user is concealing his or her emotion. However, the items of this category were not annotated often, and we decided not to use this category.

The category 'basic emotions' contains Ekman's "big six emotions" (see Ekman (1984)), with the exception of fear (unlikely in these conversations), and with the added emotion 'amusement', which would otherwise not be represented. These emotions are also represented as continuous trace annotations, which means they measure the intensity of the emotion at a given time. Epistemic states (explained by Baron-Cohen et al. (2004)) measure aspects such as how certain, agreeing, or interested the user is at a certain point of time. The category Interaction Process Analysis is a subset from the categories used by Bales (1950). Bales used these categories to classify dif-

Basic Emotions		Epistemic States		Interaction Process Analysis	
Anger	10	(not) certain	23	Shows solidarity	9
Disgust	2	(dis) agreement	79	Shows Antagonism	15
Amusement	82	(un) interested	22	Shows tension	12
Happiness	27	(not) at ease	39	Releases tension	14
Sadness	21	(not) thoughtful	41	Makes suggestion	6
Contempt	10	(not) concentrating	9	Asks for suggestion	2
				Gives Opinion	42
				Asks for opinion	3
				Gives information	72
				Asks for information	3

Table 8.1: Additional annotations dimensions (except Validity). The numbers indicate the total number of annotations for each dimension.

ferent aspects of a conversation, such as types of questions and answers, and positive or negative reactions.

All these dimensions are recorded as trace-style continuous annotations. This means that, while running a recording, the annotator can continuously move the cursor to a value between -1 and 1 to indicate the current value. For each annotation, this results in a graph with the value over time, and extracting binary labels is a matter of setting a certain threshold.

8.1.3 Automatically extracted features

We used openSMILE, the core component of OpenEAR (Eyben et al., 2009), to automatically extract features from the recordings. With openSMILE, the following low-level audio features are extracted every 10 milliseconds:

- $\mathbf{F_0}$, also called F0, is the fundamental frequency, the lowest detected frequency in the audio.
- **Intensity** is an acoustic, non-linear measure of the power of the sound.
- **Loudness** is defined as $log(intensity)$.
- **RMS-Energy** is defined as $\sqrt{\sum_{time}(intensity^2)}$.
- **LOG-Energy** is defined as $log\sum_{time}(intensity^2)$

However, a classifier needs its input in the form of a feature array with a fixed length. To create this feature array for a user's turn, the minimum, maximum, mean and standard deviation for each feature in that turn was calculated. These numbers summarize the detected features of that turn.

Besides these low-level audio features, openSMILE also detects affective and emotional states. We used the detection of valence, arousal (called activation in the annotations), potency (called power in the annotations), unpredictability (called anticipation/expectation in the annotations), intensity, and interest. These states are event-based, and are only outputted if the corresponding classifier has a result which it is confident enough about. This information is summarized by calculating the number of low (value <0), number of high (value >0), the minimum, maximum, mean and standard deviation are calculated and put into the feature set.

8.2 Sal response suggestions

The SEMAINE corpus contains 6.5 hours of human-human conversations in the same setting as the Sal, with one participant in the role of the listener. In order to use the corpus as training material to select responses for our virtual agent, we needed to map the operator utterances — which were unconstrained — to the limited list of responses that constitute the repertoire of the characters.

We took a subset of the corpus (recordings 4, 6, 9, 10, 11, 13, 15, and 16) and extracted all (893) user turns from them: that is, fragments between two operator responses where the user talks. The next step was to show these turns to human annotators, and we divided the user turns into four annotation sets. The first set contained turns 1, 5, 9 and 13, the second contained turns 2, 6, 10 and 14, and so on. With this method, the user turns had little relation to each other, which forced the annotator to look at each turn individually, without taking the context of the previous turns into account. This is similar to how we want the classifiers to work: they only take the previous user turn into account for simplicity reasons.

Each of the four annotation sets was given to three annotators. The annotators were also given the set of responses that the characters can give in the system. For each user turn in the annotation set, the annotators were instructed to select three responses in the character set that they would rate as an appropriate response to the user's utterance. Note that they were not told to select the three *best* responses, because this would result in looking through all (more than 100 per character) responses for each user turn to find the best ones. Instead, since many responses are appropriate after a user turn, the annotators were asked only to select three appropriate ones. This also increased the annotation speed, which gave them time to do more annotations.

To make the selection easier for the annotators, we grouped the responses into simple categories, such as 'starting a new topic', 'Cheering up', and 'Insults'. We told the annotators that these categories were only there to simplify the annotation process, and that it did not matter whether all three responses came from the same category or from three different ones. Also, if there were only one or two appropriate responses, that was alright too. In the end, all 893 user turns were annotated, and Table 8.2 shows details of how many user response suggestions and user turns we have.

8.3 Grouping responses

In the previous sections, we explained which features we extracted from the data and how we gathered appropriate responses. The next step was to reduce the number of classes for the classifiers by grouping the responses into categories. There are about

Number of response suggestions	Number of user turns
1	10
2	61
3	361
4	39
5	166
6	256

Table 8.2: Number of user turns that received a number (1-6) of response suggestions

480 possible responses, and we extracted only 893 user turns, with a total of 3737 response suggestions. Assuming that each response was selected an equal number of times, that means that each response was suggested about 7.8 times, which means that the classifier has on average only 7.8 training instances for each response. This is not enough to get good results.

We decided to decrease the number of classes for the classifiers by grouping the responses into categories, mainly because a lot of responses say the same thing. For the classifier this means that it only has to learn when to give a certain response *category*, while the exact response from that category can be selected randomly. Of course, the way the utterances are grouped into categories has a significant influence on the result. Assuming that similar responses occur in similar situations, grouping different responses together means that the training-situations (the features) will be different too. The classifier will then not be able to find a good pattern in the features and will not be able to make good predictions. Therefore, making good groups is essential for the quality of the results.

Responses can be grouped in many ways, there are a lot of potential methods, and since it is not clear which method is best we tried several at the same time.

- **Intuition** This grouping method was based on our own intuition of which utterances look similar, for example responses that start a new topic, try to cheer the user up or insult the user. This is the most simple way of grouping, but not necessarily the best method since it is subjective.

- **Questionnaire-based** In an attempt to remove some of the subjectivity of the previous method, we decided to group responses based on the intuition of more than one person. Using an online card-sorting study[1] we asked 12 people — three for each character — to sort the responses of that character into categories that made sense to them. After this we merged the results by grouping the responses that were often put into the same category.

- **PCA** Instead of using groups that were defined by other people (as in the previous two methods), we also grouped the responses based on the annotator's suggestions. Using a Principal Component Analysis on the suggested responses we grouped the responses that were often suggested together. We created three versions of this group, using the best 5, 10 or 20 categories.

[1] www.websort.net

- **Cluster-based** This approach bases its grouping on responses that were suggested after similar user turns. It assumes that similar user turns are followed by similar kinds of responses. From the audio fragments, the automatically extracted audio features were taken and clustered using the EM (Expectation-Maximization) or the K-Means clustering algorithms. The created clusters contain similar user turns (with similar audio features), and by grouping the suggested responses of these user turns we created the cluster-based groups.

- **Mixed** This group was created during the evaluation, by taking the best performing categories (with the highest F-score, see section 8.5.1) from the other groups, and mixing them into a new group.

- **Random** We created several random groups for comparison. In total four versions were created, two with all responses divided in three categories, and two versions with five categories. We created groups with three and five categories to verify what the effect of the number of categories is, and we created two versions of each group for better comparison (because one version could be accidentally better or worse).

8.4 Classification

This section explains which classifiers we used, and how we used them. The classifiers require data to train the models, and data to evaluate how well they perform. Therefore, we will first explain how we created the training and evaluation data.

8.4.1 Training data

To train their models, our classifiers need data, namely user turns containing features and correct responses — or in our case, response categories. Using this data the classifier tries to find patterns in the features to base its selected response on. However, there are some issues with the type of data we have. Note that we use the term *data point* for the list of features of a user turn.

One of the issues is that each user turn has at least *three* response suggestions. The most straightforward method to deal with this is to take multiple copies of each data point, one for each response suggestion. We will call this the *Full_dataset*. However, there are some problems with this approach. For example, there are multiple instances of the same data point (the exact same set of features), but potentially with different response categories, which can confuse the classifier. To solve this, we created the training data using a second method, in which only one instance of each data point and the category that was chosen most often. If more than one category was chosen most often, one of these was selected randomly. This method would ensure that each data point only occurred once, with the most common response category. This data set is called the *Single_dataset*.

However, this last method does not use all available data. For example, if a user turn is followed by two response categories, the previous method only gives the classifier one category. But if a data point has multiple response suggestions, perhaps all categories are correct. Also, a problem with the first method is that the data is potentially unbalanced: a data point could have two suggestions from category A and one suggestion from category B. When this kind of data is fed to the classifier, it probably focusses more on category A, while this is not necessarily the best approach. Therefore, a third method was used to create training data which takes all data, but removes similar categories in the same user turn. The resulting data set is called the *Unique_dataset*.

To exemplify the different methods, assume that a user turn has three response suggestions, two in category A and one in category B. The Full_dataset uses all categories. The Unique_dataset removes the double appearance of category A, and uses one user turn with category A and one user turn with Category B. The Single_dataset only uses a user turn with Category A, since this is the category that contains the most suggested responses.

8.4.2 Evaluation data

The normal method of evaluation is to make an evaluation set with data points and 'correct' answers (in our case, response categories), feed this data into the trained classifier, and match the output with the 'correct' answer. However, just as with the creation of the training data, we have the problem that one data point can have up to six possible responses. If these are all from different categories, how can we verify whether the produced response category of the classifier is correct or not? Should we match this produced category with all six suggested categories one by one? If the produced category is suggested once, this results in five incorrect matches and one correct match.

For example, say we have a user turn with three suggested responses, two from category A and one from category B. We feed this data point to the classifier, which produces category A. If we compare this answer with all three suggested categories, then it will be correct twice and incorrect once. But one could argue that a response from a good category was chosen, so the answers should be correct (or at least partially correct) each time.

We implemented several methods for this evaluation:

- **Fraction_full** This evaluator feeds each user turn to the classifier once, and evaluates the classifier's result with the fractions of the possible categories. For example, a user turn has three suggestions, two from category A and one from category B. If the classifier returns category A, this is evaluated as $\frac{2}{3}$ correct, and $\frac{1}{3}$ incorrect. This is basically the same as a binary evaluator, since evaluating the data point three times results in exactly the same proportion (two correct, one incorrect).

- **Fraction_single** A problem with the previous method is that it still evaluates a predicted category as partially incorrect if there are multiple possible categories. For example, in the given example, even though category A was suggested twice by an annotator and also predicted by the classifier, it is still evaluated as $\frac{1}{3}$ incorrect. To rectify this issue, the fraction_single evaluator does not evaluate a predicted category as incorrect if it is partially correct. So, in the previous example, this would be evaluated as $\frac{2}{3}$ correct only.

- **Fraction_single_recoded** The fraction_single evaluator creates unbalanced results because an incorrect result is evaluated as fully incorrect, but a correct result could be evaluated as (for example) $\frac{2}{3}$ or $\frac{1}{3}$ correct. To solve this problem the fractions are recoded. This recoding is based on the distribution of the suggested categories, and each fraction is divided by the fraction that occurs most often. For example, in the on-going example the possible categories are [AAB], which means that A occurs two out of three times and B one out of three times. By dividing these fractions by the highest fraction (in this case $\frac{2}{3}$), the new recoded fractions are 1 for category A, and $\frac{1}{2}$ for category B. In this case, if the classifier returns category A, this is evaluated as completely correct. If the classifier returns category B, then this is only $\frac{1}{2}$ correct, because category A is suggested twice as often as B.

- **Exists** An even more lenient method is an evaluator that always evaluates a predicted category as fully correct if it was annotated at least once. In our example, this means that if the classifier returns category A or B, this is evaluated as fully correct in both cases because both category A and B have been suggested at least once. A downside of this method is that if a lot of different categories are suggested for a particular user turn, then the classifier has a very high (perhaps unreasonable) chance of predicting a suggested category.

In the end, it is hard to tell which of these measures is a 'fair' evaluation method that returns representative results (so a meaningful number between 0 and 1). The methods given are ordered from very strict (resulting in lower values) to very lenient (resulting in higher values). However, when using the evaluation to identify relative differences between approaches, then the evaluation method does not really matter, as long as the same evaluation method is used in each approach.

8.4.3 Classifiers

The goal of these classifiers is of course to create working models for response selection, but a second important goal is to learn how these models work. Therefore, we decided to only use machine learning techniques that produce human-readable and understandable models. The first classifier is the decision tree algorithm J48, which is a Java implementation of the C4.5 algorithm (Salzberg, 1994). This classifier produces a decision tree, a tree-like model. The nodes in the tree represent the 'choices' to make, usually a comparison of a value which determines which path to take from there. The leaves are labelled with the predicted response. The second classifier we used is Ripper (Cohen, 1995), an algorithm that produces simple rules in the form:

```
if value1 < v1 and value2 > v2 and value3 < v3 => response A
```

8.5 Performance results

This section presents the results of the response classification. Using each possible configuration of the different trainers and evaluators, we passed the data through both classifiers using ten-fold cross-validation. In the first round we used the described trainers and evaluators, and based on the results we made some improvements.

8.5.1 First round

Table 8.3 shows the initial classifier results, with for each group and evaluation method the average F-value (harmonic mean of precision and recall) of the three training methods and two classifiers.

	# of	Fraction_full					Fraction_single					Fraction_recoded					Exists				
Group	categories	Ob	Po	Pru	Spi	**Avg**	Ob	Po	Pru	Spi	**Avg**	Ob	Po	Pru	Spi	**Avg**	Ob	Po	Pru	Spi	**Avg**
Cluster-based	3.25	0.19	0.28	0.19	0.32	**0.24**	0.29	0.40	0.30	0.48	**0.37**	0.38	0.46	0.37	0.54	**0.44**	0.41	0.50	0.40	0.57	**0.47**
Intuition	6.75	0.15	0.17	0.14	0.12	**0.14**	0.18	0.21	0.17	0.14	**0.17**	0.23	0.25	0.22	0.19	**0.22**	0.25	0.27	0.25	0.23	**0.25**
PCA_20	20	0.03	0.04	0.03	0.04	**0.03**	0.04	0.05	0.04	0.05	**0.04**	0.07	0.10	0.06	0.09	**0.08**	0.09	0.11	0.08	0.11	**0.09**
PCA_10	10	0.06	0.08	0.06	0.07	**0.07**	0.09	0.12	0.09	0.10	**0.10**	0.11	0.17	0.11	0.15	**0.13**	0.11	0.18	0.11	0.18	**0.15**
PCA_5	5	0.13	0.14	0.14	0.13	**0.13**	0.18	0.20	0.19	0.18	**0.19**	0.20	0.23	0.22	0.21	**0.21**	0.21	0.25	0.24	0.21	**0.23**
Question_Full	15	0.05	0.05	0.05	0.05	**0.05**	0.06	0.07	0.06	0.07	**0.06**	0.10	0.13	0.10	0.11	**0.11**	0.10	0.15	0.12	0.13	**0.12**
Question_Good	8.5	0.09	0.10	0.08	0.10	**0.09**	0.12	0.14	0.11	0.15	**0.13**	0.19	0.20	0.17	0.19	**0.19**	0.21	0.22	0.19	0.21	**0.21**
Random3A	3	0.21	0.19	0.18	0.20	**0.19**	0.35	0.34	0.29	0.35	**0.33**	0.45	0.41	0.35	0.43	**0.41**	0.49	0.43	0.37	0.45	**0.44**
Random3B	3	0.20	0.19	0.21	0.18	**0.19**	0.35	0.33	0.38	0.32	**0.35**	0.42	0.39	0.48	0.39	**0.42**	0.45	0.42	0.50	0.41	**0.44**
Random5A	5	0.12	0.13	0.12	0.11	**0.12**	0.20	0.21	0.21	0.18	**0.20**	0.30	0.33	0.32	0.28	**0.31**	0.34	0.37	0.36	0.32	**0.35**
Random5B	5	0.12	0.12	0.11	0.14	**0.12**	0.19	0.19	0.18	0.23	**0.20**	0.26	0.29	0.25	0.36	**0.29**	0.30	0.33	0.30	0.39	**0.33**

Table 8.3: Initial classifier results, showing the F-value (harmonic mean of precision and recall) of each group and evaluation method. Each value is the average of the three training methods and the two classifiers.

This table shows that the evaluators are indeed ordered by leniency, where the leftmost evaluator is the most strict and results in the lowest values, and the rightmost evaluator is the most lenient with the highest values. It also shows that the cluster-based groups performed best, together with the random groups with three categories (Random3A and Random3B). This can be explained by the fact that these groups contain fewer categories, which means that the chance that the classifier returns the correct category is higher. We will go into more detail later, but first we will show the differences between the trainers and classifiers used.

Table 8.4 and 8.5 show the differences between the classifiers and training methods used.

	Fraction_full					Fraction_single					Fraction_recoded					Exists				
Classifier	Ob	Po	Pru	Spi	**Avg**	Ob	Po	Pru	Spi	**Avg**	Ob	Po	Pru	Spi	**Avg**	Ob	Po	Pru	Spi	**Avg**
J48	0.16	0.18	0.15	0.17	**0.16**	0.24	0.27	0.23	0.26	**0.25**	0.32	0.35	0.31	0.33	**0.33**	0.35	0.38	0.35	0.37	**0.36**
JRIP	0.12	0.14	0.11	0.13	**0.12**	0.18	0.20	0.16	0.18	**0.18**	0.22	0.24	0.20	0.22	**0.22**	0.23	0.25	0.21	0.24	**0.23**

Table 8.4: The differences between the two classifiers. The values are averages of all trainers and evaluators.

	Fraction_full					Fraction_single					Fraction_recoded					Exists				
Trainer	Ob	Po	Pru	Spi	**Avg**	Ob	Po	Pru	Spi	**Avg**	Ob	Po	Pru	Spi	**Avg**	Ob	Po	Pru	Spi	**Avg**
Full	0.15	0.16	0.14	0.16	**0.15**	0.22	0.25	0.21	0.24	**0.23**	0.28	0.31	0.28	0.30	**0.29**	0.31	0.33	0.31	0.33	**0.32**
Single	0.15	0.17	0.14	0.16	**0.15**	0.23	0.26	0.21	0.24	**0.23**	0.30	0.33	0.27	0.31	**0.31**	0.33	0.38	0.30	0.34	**0.34**
Unique	0.12	0.14	0.12	0.13	**0.13**	0.17	0.19	0.17	0.18	**0.18**	0.22	0.23	0.21	0.22	**0.22**	0.23	0.24	0.23	0.24	**0.24**

Table 8.5: The differences between the three training methods. The values are averages of all classifiers and evaluators.

From these tables, it looks as if the J48 algorithm performs best, and the Full and Single training method too. However, we decided to dig a bit deeper for details. Table 8.6 shows the results of different combinations of classifiers and trainers. Because this table is meant to demonstrate the relative differences, the values are based on only one evaluator, namely Fraction_full. This table shows that the J48 algorithm works best in combination with the Full and the Single trainers. The Unique trainer performs worse, mostly in combination with the JRip classifier (see the bottom row in Table 8.6).

Classifier	Trainer	Cluster-based	Random3A	Random3B	Random5A	Random5B	PCA_5	Intuition	Ques_Good	PCA_10	Ques_Full	PCA_20
J48	Full	0.28	0.24	0.26	0.18	0.17	0.13	0.16	0.13	0.08	0.07	0.05
J48	Single	0.27	0.26	0.24	0.15	0.16	0.17	0.17	0.13	0.08	0.06	0.05
J48	Unique	0.26	0.16	0.18	0.16	0.14	0.12	0.18	0.11	0.07	0.06	0.05
JRip	Full	0.25	0.19	0.17	0.09	0.08	0.12	0.15	0.08	0.05	0.04	0.02
JRip	Single	0.25	0.18	0.18	0.09	0.11	0.14	0.14	0.07	0.07	0.04	0.02
JRip	Unique	0.15	0.15	0.15	0.07	0.09	0.12	0.08	0.05	0.05	0.02	0.02

Table 8.6: The results of different combinations of classifiers and trainers. The evaluator Fraction_full was used for the results.

Some other conclusions could be drawn based on the results so far. To start, some groups performed badly, especially the ones with a lot of categories. This makes sense for several reasons. Firstly, if there are more categories to choose from, the chance of selecting the correct one is lower. Secondly, more categories means less training data per category, lowering the quality of the classifier. Lastly, if there are more categories the chance that some categories will overlap — they can be selected in certain similar contexts — increases. This means that the chance that an appropriate category will be predicted by the classifier which has not been annotated increases as well.

Another look at Table 8.3 shows that the random groups perform relatively well. This is tied to the previous point, since the random groups only contain three or five categories, while other groups sometimes contain more than 10 categories. To make comparisons that are more valid, in the next section we will modify our random groups.

We will not show the details here, but we also have access to the results of each category, showing exactly which category of which group performed well and which did not. We can use this information to improve our models by removing the categories that did not perform well. This might seem strange at first, because this effectively means that the classifier can predict less. However, our goal is to create response models that can predict a certain response in certain contexts, but it does not have to predict a response *at all times*. If the classifier is not sure at a certain moment, it is best not to use its output but to use a different model instead. So, we are mainly looking for categories that can be (more or less) reliably predicted.

8.5.2 Different feature sets

In section 8.1 we explained that we used two types of features: the automatically extracted audio features and the human-made annotations. The models in the previous section only use the audio features, mostly because the final models should be able to work in a real system. But what would happen if the models were trained and evaluated on the annotations? Since the annotations are man-made, one could expect the quality to be higher than automatically extracted features. But would the quality of the models improve too? Also, the annotations contain different information than the extracted features, and we would like to know how this information affects the classifiers.

This section explores the differences between the extracted features and the annotations. The experiment from the previous section was performed again, with some small variations. Only the groups Intuition, PCA_5, Questionnaire_full, Questionnaire_good, and cluster-based were used, and only the Full and Single trainer. We removed the groups PCA_10 and PCA_20 and the Unique trainer since these performed worse than the others.

Using these groups and trainers we trained and evaluated the classifiers again, but this time with different parts of the annotations. Since all recordings were annotated with the five core dimensions — Valence, Activation, Power, Anticipation/Expectation and Intensity — we used these five dimensions in all tests. Additionally, we tried adding the dimensions that were annotated most often, namely agreement, gives information, amusement, and gives opinion. The results of these dimensions can be found in Table 8.7.

	Fraction_full					Fraction_single					Fraction_recoded					Exists				
Category	Cluster-based	Intuition	PCA_5	Questionnaire_Full	Questionnaire_Good	Cluster-based	Intuition	PCA_5	Questionnaire_Full	Questionnaire_Good	Cluster-based	Intuition	PCA_5	Questionnaire_Full	Questionnaire_Good	Cluster-based	Intuition	PCA_5	Questionnaire_Full	Questionnaire_Good
Extracted audio-features	0.24	0.14	0.13	0.05	0.09	0.37	0.17	0.19	0.06	0.13	0.44	0.22	0.21	0.11	0.19	0.47	0.25	0.23	0.12	0.21
Core-dimensions	0.22	0.14	0.13	0.05	0.09	0.33	0.17	0.19	0.07	0.13	0.42	0.22	0.22	0.11	0.19	0.47	0.25	0.24	0.12	0.22
Core+Agreement	0.22	0.14	0.13	0.05	0.09	0.35	0.17	0.19	0.07	0.13	0.44	0.22	0.22	0.11	0.19	0.50	0.24	0.25	0.13	0.23
Core+Amusement	0.23	0.13	0.13	0.04	0.09	0.36	0.16	0.21	0.06	0.13	0.46	0.21	0.24	0.09	0.19	0.52	0.24	0.28	0.11	0.22
Core+Information	0.23	0.15	0.13	0.05	0.10	0.35	0.19	0.20	0.07	0.15	0.43	0.24	0.24	0.11	0.21	0.47	0.26	0.26	0.13	0.25
Core+Opinion	0.21	0.14	0.13	0.05	0.09	0.32	0.17	0.20	0.06	0.13	0.40	0.22	0.23	0.11	0.19	0.44	0.25	0.26	0.12	0.22

Table 8.7: A comparison between different feature sets. The numbers are the average values of all four characters, using both the J48 and JRip classifiers with the Full and the Single trainers.

This table shows that the classifiers performed very similarly with different features. The classifier with the extracted audio features had an average score of 0.16, which is the same average score as the classifiers that use the annotations. Only the classifier that used the Core and the Information annotations scored slightly better with an average score of 0.18.

It is possible that the annotations will not provide more useful information for response selection than features that can be automatically extracted. Or perhaps the quality of the annotations was too low to be useful. It could be one of these reasons, or perhaps both. In any case, in the next sections we will only use the extracted audio-features, since we could use the classifiers that are created with this data in an online system as well.

8.5.3 Validating the cluster-based method

The cluster-based group — created by grouping responses that were suggested after similar user turns — performed better than most other groups, but there could be a very good reason for this: the grouping was based on the same data that was used for training and evaluation. The categories consisted of responses that followed similar user turns, which increases the likelihood that the classifier will find the same patterns and will be able to distinguish these categories more easily.

To verify whether this is a problem or not, we checked if a group that is created from different data performs similar to a group that is created from all data. We split the data in two equal-sized parts (imaginatively called 'part1' and 'part2'), and performed the same grouping-strategy — using EM or K-means algorithms to cluster the user turns — to create a group for each part. Then we ran the classifiers over the same parts, using both groups and the original cluster-based grouping. We expected to see that the groups of the same data would perform better than the other group, but what we wanted to see was that the other group did not perform too badly. We used only the J48 algorithm, and both the Full and Single trainers. The results can be found in Table 8.8.

	Fraction_full					Fraction_single					Fraction_recoded					Exists				
Category	Ob	Po	Pru	Spi	**Avg**	Ob	Po	Pru	Spi	**Avg**	Ob	Po	Pru	Spi	**Avg**	Ob	Po	Pru	Spi	**Avg**
	Using the part1 data																			
Cluster-based	0.23	0.33	0.21	0.33	**0.27**	0.29	0.42	0.25	0.40	**0.34**	0.37	0.47	0.31	0.46	**0.40**	0.41	0.51	0.32	0.49	**0.43**
Clustered_part1	0.36	0.34	0.30	0.45	**0.36**	0.44	0.42	0.39	0.55	**0.45**	0.54	0.50	0.48	0.62	**0.53**	0.57	0.52	0.50	0.64	**0.55**
Clustered_part2	0.26	0.32	0.25	0.28	**0.28**	0.33	0.39	0.30	0.33	**0.34**	0.43	0.47	0.39	0.41	**0.42**	0.46	0.49	0.42	0.43	**0.45**
	Using the part2 data																			
Cluster-based	0.22	0.33	0.24	0.35	**0.29**	0.28	0.41	0.29	0.42	**0.35**	0.36	0.45	0.37	0.46	**0.41**	0.39	0.48	0.41	0.49	**0.44**
Clustered_part1	0.25	0.29	0.26	0.36	**0.29**	0.32	0.35	0.34	0.44	**0.36**	0.43	0.43	0.42	0.52	**0.45**	0.46	0.46	0.44	0.55	**0.48**
Clustered_part2	0.33	0.39	0.31	0.32	**0.34**	0.42	0.46	0.39	0.38	**0.41**	0.52	0.52	0.45	0.46	**0.49**	0.54	0.55	0.48	0.48	**0.51**

Table 8.8: Results of splitting the data in two parts, creating cluster-based groups of each part and train and evaluate the classifiers with this data.

This table shows that the groups that were tested on different data than the ones they were created from had the same performance as the original cluster-based groups. This shows that this approach is also valid when different data is used.

8.5.4 Improving the models

Based on the results of the first round we could improve the models, because some categories performed really badly. The intuition and the questionnaire-based models were improved by removing these categories. The PCA_10 and PCA_20 groups were removed completely due to their low performance.

To make a fairer comparison, we created a random version of each existing group, with the exact same number of categories, and each category containing the exact same number or responses. This should have ensured a fair comparison with a random model. The results of these new models can be found in Table 8.9.

This table shows that most models were indeed better than randomly selecting a group. The PCA_5 group is the only one that performed worse than random with each evaluator. The Intuition group performed worse than random with more lenient evaluators (fraction_recoded and exists). Appearantly, that particular random-group had its responses divided into categories in such a way that suggested responses easily fell into multiple categories. This means that a user turn had multiple suggested categories, which raised its score with the more lenient evaluators.

	# of	Fraction_full					Fraction_single					Fraction_recoded					Exists				
Group	categories	Ob	Po	Pru	Spi	**Avg**	Ob	Po	Pru	Spi	**Avg**	Ob	Po	Pru	Spi	**Avg**	Ob	Po	Pru	Spi	**Avg**
Cluster-based	3.25	0.26	0.30	0.21	0.34	**0.28**	0.41	0.45	0.32	0.52	**0.42**	0.54	0.53	0.43	0.59	**0.52**	0.60	0.60	0.50	0.63	**0.58**
Cluster-based_Rand	3.25	0.25	0.24	0.16	0.25	**0.22**	0.43	0.25	0.27	0.41	**0.34**	0.60	0.25	0.33	0.51	**0.42**	0.66	0.25	0.38	0.57	**0.46**
Intuition_Impr	4.5	0.19	0.27	0.21	0.30	**0.24**	0.24	0.32	0.25	0.37	**0.29**	0.30	0.38	0.35	0.45	**0.37**	0.34	0.42	0.41	0.51	**0.42**
Intuition_Impr_Rand	4.5	0.15	0.19	0.16	0.19	**0.17**	0.24	0.31	0.25	0.33	**0.28**	0.38	0.46	0.38	0.45	**0.42**	0.45	0.55	0.45	0.48	**0.48**
Mixed	2.25	0.35	0.39	0.32	0.33	**0.35**	0.49	0.50	0.41	0.46	**0.46**	0.57	0.53	0.49	0.52	**0.53**	0.63	0.59	0.54	0.60	**0.59**
Mixed_Rand	2.25	0.32	0.30	0.23	0.32	**0.29**	0.44	0.42	0.38	0.42	**0.41**	0.45	0.46	0.48	0.44	**0.46**	0.47	0.55	0.56	0.45	**0.51**
PCA_5	5	0.14	0.16	0.15	0.15	**0.15**	0.22	0.25	0.23	0.21	**0.23**	0.26	0.33	0.31	0.27	**0.29**	0.29	0.39	0.36	0.28	**0.33**
PCA_5_Random	5	0.16	0.18	0.16	0.16	**0.16**	0.26	0.29	0.26	0.27	**0.27**	0.41	0.45	0.42	0.42	**0.42**	0.46	0.50	0.48	0.49	**0.48**
Questionnaire_Impr	5.25	0.19	0.21	0.15	0.21	**0.19**	0.26	0.29	0.21	0.30	**0.26**	0.38	0.39	0.31	0.35	**0.36**	0.43	0.45	0.38	0.41	**0.42**
Questionnaire_Impr_Rand	5.25	0.13	0.14	0.12	0.16	**0.14**	0.21	0.19	0.19	0.22	**0.20**	0.32	0.23	0.26	0.25	**0.26**	0.38	0.27	0.31	0.29	**0.31**

Table 8.9: The improved classifier results, with a random version of each tested group for fair comparison.

Unfortunately, it is not possible to calculate the exact performance of each classifier. Part of the reason for this is the data: the fact that each user turn has multiple responses makes it hard to evaluate properly. But another, major reason is the fact that during the annotation, the raters did not select *all* appropriate responses but only three. This means that we do not know exactly which responses are appropriate and which are not. If an appropriate response is predicted by the classifier, but was not suggested as such by the annotators — because he selected three other responses which were also appropriate — then the response is usually evaluated as inappropriate, even though it is appropriate.

The evaluations performed in this section show that the models can at least roughly predict appropriate responses, but in order to evaluate how well they really perform a different kind of evaluation is needed. In the next section we will explain how we performed this evaluation by showing the user turns and the predicted responses to users, so that they could directly rate that response.

8.6 Online evaluation

In order to evaluate whether our classifiers can be used to select responses for new user turns, we decided to perform an evaluation in which humans had to rate the responses suggested by those classifiers. Using new fragments from the SEMAINE corpus (recording 17, 18 and 19), we extracted 50 fragments containing a user turn, with an average length of 12.94 seconds (SD 8.97 seconds). We used openSMILE to extract audio features from these fragments, and we used our classifiers to produce possible responses for each user turn.

We used our best-performing models. For the groups, we used the cluster-based, the questionnaire-improved, the intuition-improved, and the mixed group. We took the J48 algorithm and the Full trainer, since these performed best. These classifiers predicted a response category for each user turn, together with a confidence value. Also, all responses that were not included in the categories were put in the category 'None', which was used to indicate that the classifier does not know an appropriate response. If another category was selected by the classifier but the confidence value was below 0.5, then it was converted to None as well. This meant that for every user turn, each classifier could return a good response category, but it was also possible for a classifier to return nothing. This resulted in one to four categories per user turn. For each category, three responses were randomly selected from that category, giving the user turn three to twelve actual responses. Additionally, we added two completely random (selected from all possible responses) responses, resulting in five to fourteen possible responses per user turn.

In short, we used four different classifiers, and each could suggest a category or not. From each category, three responses were randomly selected. Additionally, we added two random responses to the list.

In the end, we had 50 short clips, and for each clip four different response lists based on a certain SEMAINE character, containing five to fourteen possible responses for each character. Ten clip-character combinations were taken out because no classifier could select an appropriate response, resulting in 190 combinations. We put these clips in an online questionnaire, in which we asked participants to view 50 clips and rate the quality of the classifier's responses The clips were shown in a random order, as were the classifier's responses. Each response was rated on a 7-point Likert scale. In total, 21 people participated, resulting in a total number of ratings of 613 user turns and a total of 4841 responses.

To calculate the performance of the different models, we calculated the average rating for each character and category combination. These results can be found in Table 8.10.

This table shows the average ratings, the standard deviations, and the total number of responses that were annotated. To find out which models performed significantly better or worse than others, the next step was to perform an ANOVA test, followed by a Games-Howell post-hoc test, chosen because of unequal sample sizes. The results, combined with a visual representation of the mean ratings, can be found in Figure 8.3.

These results show that for Poppy and Prudence, the classifiers performed better than the random model, although unfortunately only the Mixed model and the Questionnaire model for Poppy were significantly better. For Obadiah, the random model performed similarly to the classifiers, and for Spike the random model performed even better than the classifiers, in one case even significantly better.

However, there are some minor issues with the clips and the responses that possibly skewed the results. There could have been some apparently ill-chosen video clips that were very hard to respond to appropriately with the limited set of responses we had. Also, there could have been responses that were simply not appropriate at all, except on some very rare occasions. To filter these issues out, we selected three clips that performed too low (an average score below 3) and removed them from the data. To find inappropriate responses, we selected responses that were rated at least 6

	Cluster-based			Intuition			Mixed			Questionnaire			Random		
	Mean	SD	N	**Mean**	SD	N	**Mean**	SD	N	**Mean**	SD	N	**Mean**	SD	N
Poppy	**3.87**	1.94	315	**3.84**	2.03	399	**4.32**	1.73	60	**4.4**	1.7	57	**3.48**	2	312
Prudence	**3.8**	1.79	183	**3.81**	1.86	285	**3.69**	1.92	219	**4**	1.99	204	**3.61**	1.9	294
Obadiah	**4.02**	2.03	375	**3.7**	2.01	480	**3.27**	1.96	123	**3.7**	2.1	249	**3.74**	2.03	320
Spike	**3.66**	2.08	369	**3.44**	2.07	126	**0**	0	0	**3.37**	1.98	171	**4.02**	2.00	300

Table 8.10: Average ratings for the different groups and characters.

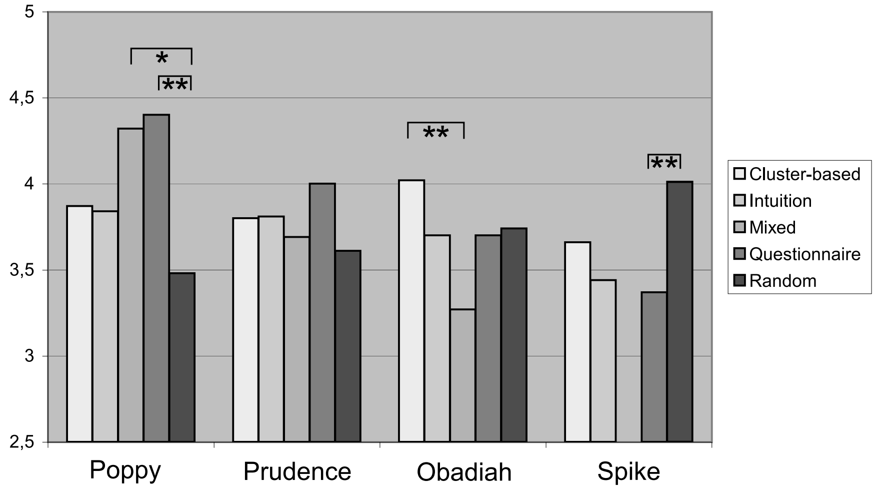

Figure 8.3: Average ratings for the different groups and characters, and significant differences. * = p <0.05, ** = p <0.01, *** = p <0.005

times, with an average rating below 1.5. Five responses fulfilled this requirement, and they were removed from the data. Note that these removals are not biased towards a single classifier model, they only remove clips and responses that under-perform. Detailed results can be found in Table 8.11, and a visual representation of the average ratings and their significant differences can be found in Figure 8.4.

	Cluster-based			Intuition			Mixed			Questionnaire			Random		
	Mean	SD	N	**Mean**	SD	N	**Mean**	SD	N	**Mean**	SD	N	**Mean**	SD	N
Poppy	**3.94**	1.92	297	**3.87**	2.01	378	**4.32**	1.73	60	**4.4**	1.7	57	**3.55**	2	297
Prudence	**3.8**	1.79	183	**3.83**	1.86	276	**3.87**	1.9	189	**4.07**	1.99	195	**3.51**	1.86	268
Obadiah	**4.21**	1.98	336	**3.82**	1.99	434	**3.6**	1.93	93	**3.92**	2.07	219	**3.9**	2	288
Spike	**3.69**	2.07	351	**3.72**	1.97	105	**0**	0	0	**3.44**	1.91	153	**4.1**	1.96	282

Table 8.11: Average ratings for the different groups and characters, based on the filtered results.

These results show some improvements. The four classifiers of Prudence are now rated even higher than the random model, and the Questionnaire model is now significantly better too. The models for Obadiah and Spike have not improved. This means that for Poppy and Prudence, some of our classifiers produce responses that are significantly more appropriate than random responses, with an average rating greater than four on a seven point scale.

8.7 Discussion

In this chapter we have explained how we trained classifiers to select the next response of the agent, based on the previous turn of the user. We evaluated these classifiers in two ways, namely by calculating the classifiers' performance and by having humans rate their suggested responses. We have shown, when calculating the performance of the classifiers, that they performed better than a random model. At first sight, the differences are not that great, but when calculating the improvement the classifier performed up to 41% better than the random model (using the most strict

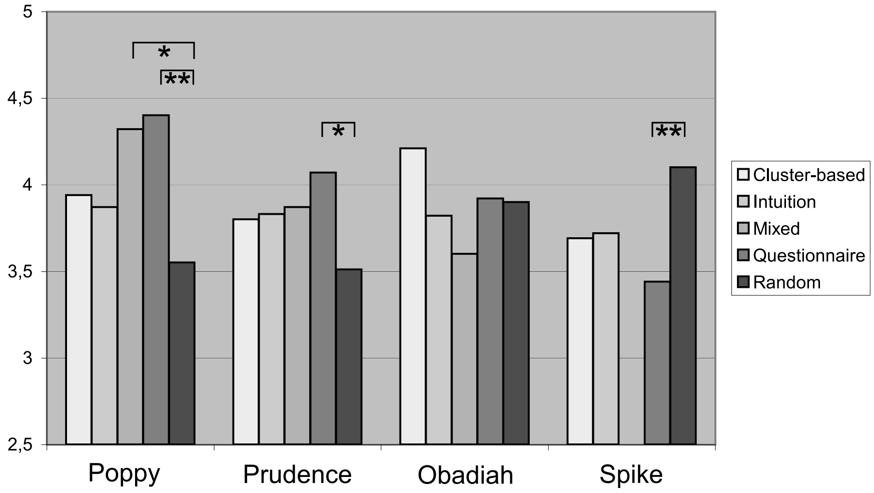

Figure 8.4: Average ratings for the different groups and characters, and significant differences, based on the filtered results. * = p <0.05, ** = p <0.01, *** = p <0.005

evaluation method). But even so, the performance results are still far from perfect, and there are several reasons for this.

The first cause that the results are still far from perfect is the structure of the training and evaluation data, in which each user turn can have up to six response suggestions. Also, the fact that it has multiple response suggestions does not mean that the other responses are not appropriate. These issues cause problems when training the classifiers, although we have tried to minimize the problems by using the data in different ways.

- The FULL dataset contains all data, but also contains multiple response suggestions for each user turn which are all correct. Also, when multiple responses belong in the same category, this potentially creates an imbalance in the data by introducing user turn–response category combinations that are equal to each other (for more details, see section 8.4.1).

- The UNIQUE dataset removes these double combinations, but this still leaves the problems of multiple possible responses for each user turn.

- The SINGLE dataset contains only one response suggestion for each user turn by choosing the response that was selected most often. The downside of this method is that it throws away a lot of data.

These methods try to deal with the problems of the data, but they all have their limitations. And the same problems emerge in the evaluation of the results, since if there are multiple 'correct' responses, which one should be used to check the result? Again, we have developed different methods that vary the strictness of the evaluation. For more details of the different evaluation methods, see section 8.4.2. Suffice to say that there still is no 'best' way to evaluate these classifiers and to get 'true' accuracy values.

A second reason for the imperfect performance has to do with the grouping of the responses. We have shown that the grouping is very important, and creating good groups is necessary to get classifiers that perform well. But as we have shown in the first reason, grouping responses can lead to similar combinations of user turn and response suggestion, which confuses the training of the classifiers. Also, grouping assumes that responses in a category are interchangeable, that is, in every situation you can use any response from the same group. We made this assumption for simplicity's sake, but it probably does not work this way. It is more likely that responses belong in different and multiple groups, and that depending on the context the grouping changes.

A third cause is the set of input features. We only used simple audio features and some affective dimensions (derived from the audio too), since the models should work in a speech-only system as well. Using more features could improve the classifiers a lot. For example, the video features could be used, more specifically the facial expressions and the head movements of the user. Additionally, more complex interpreted features could be used, such as boredom, agreement, the start of a new topic, and so on.

The online evaluation shows that for Poppy and Prudence most classifiers produced better results than a random model. Three models even produced responses with an average rating higher than four (on a seven point scale). These results demonstrate that the classifiers can use simple audio features to suggest appropriate responses. However, only the classifiers for Poppy and Produce performed well; the models for Obadiah performed similar to random, and Spike's models performed even worse than random. Why?

We argue that Poppy and Prudence are easier because they are more positive. Obadiah, the depressed character, uses most of its responses to whine about how awful life is and that the situation is completely hopeless. It is likely that the human raters did not find these responses appropriate in a lot of situations. The same holds for Spike, whose only goal is to aggravate the user, and a lot of its responses are insults to the user. Again, it is very likely that humans did not find this appropriate behaviour. However, this does make us wonder whether the raters evaluated the responses as the appropriate character. So, did they rate most of Spike's responses low because these responses were not appropriate for Spike in the conversation, or because they were not appropriate in the conversation *independent* of the current character? We do not know, but we do know that we need to make the context (that is, the current character) more clear in future evaluations.

8.8 Conclusions and future work

In the previous chapter and this one we showed different approaches to appropriate response selection for our virtual listening agent. In the previous chapter we looked at other dialogue systems, specifically how they selected responses and how that is tied to the type of dialogue system and the domain it is operating in. We looked at task-oriented systems, question-answer systems, and dialogue systems that operate in small domains. We continued by explaining that Sal is different and that the usual approaches cannot be applied directly. Sal does not take the content of the conversation into account. As a listening agent its role is only to react to the user and motivate him or her to continue speaking, and Sal can only select a response from a small finite list. We completed the chapter by discussing an initial attempt to hand-craft dialogue models, which react to detected events in the conversation such as a high or low arousal of the user, a longer period of silence, or a user's smile.

In this, we explained how we used a more data-driven approach to create dialogue models. Using videos of human-human conversations, we extracted turns from which we automatically extracted audio features. Human annotators then had to suggest three possible responses that an agent could give after each extracted turn. To lower the total number of possible responses, we used several methods of grouping them into response categories. These extracted audio features, the suggested responses, and the different groups were then fed into several classifiers to train them to give appropriate responses after a user's turn.

We evaluated the classifiers by matching their output with the suggested responses, which showed that most models performed better than random models, even up to 40% better. We continued with an online evaluation, in which we used the trained classifiers to predict appropriate responses for new turns. The online evaluation contained several video clips with a user's turn, and below that a randomized list with all predicted responses and two random responses. The evaluators had to rate how appropriate each response was after the given user's turn. This showed that all models for Poppy and Prudence performed better than the random model, of which three models significantly better with an average rating of four (on a seven point scale). This shows that even with simple audio features it is possible to create decent classifiers that produce appropriate responses based only on the previous turn of the user.

But of course there are a lot more things to do and to study. For example, it would be worthwhile to use more types of features, such as facial expressions and head movements, more higher-level features such as agreement and boredom, and perhaps even some features derived from the actual content of the user's turn.

Another point of interest is the classifier itself. We used two techniques that produce human-readable models, mostly because we wanted to see what we could learn from the rules. It would be interesting to see whether other black box classifiers possibly perform better. Also, we did not tune specific parameters of the classifiers, so there is another chance to possibly increase the performance.

It would also be interesting to see what the effect is of the limited response annotations we had. A lot of work-arounds had to be created to fix the problems that were introduced by having one to six possible suggestions. By annotating *all* possible responses for each user turn, the classifiers would know exactly which responses are appropriate and which are not. It would be interesting to see whether these classifiers would perform better than ours, and what the differences would be. A downside of this method is that it would require a lot of work to annotate the complete dataset, and when multiple annotators are needed to do this a good method to fix disagreements has to be found.

Bibliography

Arya, A., DiPaola, S., and Parush, A. (2009). Perceptually Valid Facial Expressions for Character-Based Applications. *International Journal of Computer Games Technology*, 2009.

Babu, S., Schmugge, S., Barnes, T., and Hodges, L. (2006). ”What Would You Like to Talk About?” An Evaluation of Social Conversations with a Virtual Receptionist. In Gratch, J., Young, M., Aylett, R., Ballin, D., and Olivier, P., editors, *Proceedings of the 6th international conference on Intelligent Virtual Agents*, volume 4133 of *Lecture Notes in Computer Science*, pages 169–180. Springer Berlin / Heidelberg.

Babu, S., Schmugge, S., Inugala, R., Rao, S., Barnes, T., and Hodges, L. F. (2005). Marve: a prototype virtual human interface framework for studying human-virtual human interaction. In *Proceedings of the 5th international conference on Intelligent Virtual Agents*, pages 120–133, London, UK. Springer-Verlag.

Baggia, P., Burkhardt, F., Pelachaud, C., Peter, C., and Zovato, E. (2009). Emotion Markup Language (EmotionML) 1.0. World Wide Web Consortium, http://www.w3.org/TR/emotionml.

Bales, R. F. (1950). *Interaction process analysis: A method for the study of small groups*. Addison-Wesley.

Balomenos, T., Raouzaiou, A., Karpouzis, K., Kollias, S., and Cowie, R. (2003). An Introduction to Emotionally Rich Man-Machine Intelligent Systems. In *Third European Symposium on Intelligent Technologies, Hybrid Systems and their implementation on Smart Adaptive Systems*.

Barkhuysen, P., Krahmer, E., and Swerts, M. (2008). The interplay between the auditory and visual modality for end-of-utterance detection. 123(1):354–365.

Baron-Cohen, S., Golan, O., Wheelwright, S., and Hill, J. J. (2004). Mind Reading: The Interactive Guide to Emotions.

Beattie, G. (1981). Interruption in Conversational Interaction, and its Relation to the Sex and Status of the Interactants. *Linguistics. An Interdisciplinary Journal of the Language Sciences La Haye*, 19(1-2):15–35.

Bernsen, N. O., Charfuelàn, M., Corradini, A., Dybkjær, L., Hansen, T., Kiilerich, S., Kolodnytsky, M., Kupkin, D., and Mehta, M. (2004). First prototype of conversational H.C. Andersen. In *Proceedings of the working conference on Advanced visual interfaces*, AVI ’04, pages 458–461, New York, NY, USA. ACM.

Bernsen, N. O. and Dybkjær, L. (2004). Domain-Oriented Conversation with H.C. Andersen. In André, E., Dybkjær, L., Minker, W., and Heisterkamp, P., editors, *Affective Dialogue Systems*, volume 3068 of *Lecture Notes in Computer Science*, pages 142–153. Springer Berlin / Heidelberg.

Bevacqua, E., De Sevin, E., Pelachaud, C., McRorie, M., and Sneddon, I. (2010). Building credible agents: Behaviour influenced by personality and emotional traits. In *Proceedings of the International Conference on Kansei Engineering and Emotion Research (KEER'10)*, pages 1071–1080.

Bevacqua, E., Mancini, M., and Pelachaud, C. (2008). A Listening Agent Exhibiting Variable Behaviour. *Journal of Pragmatics*, 5208:262–269.

Boersma, P. and Weenink, D. (2001). Praat, a System for Doing Phonetics by Computer. *Glot International*, 5(9/10):341–345.

Bosch, K., Harbers, M., Heuvelink, A., and Doesburg, W. (2009). Intelligent Agents for Training On-Board Fire Fighting. In Duffy, V. G., editor, *Proceedings of the 2nd International Conference on Digital Human Modeling*, volume 5620 of *Lecture Notes in Computer Science*, pages 463–472, Berlin, Heidelberg. Springer Berlin Heidelberg.

Brown, P. and Levinson, S. C. (1987). *Politeness: Some universals in language use*. Cambridge University Press.

Bui, T. (2006). Multimodal dialogue management-state of the art. Technical Report TR-CTIT-06-01, Centre for Telematics and Information Technology, University of Twente, Enschede.

Burnett, D. C., Walker, M. R., and Hunt, A. (2004). Speech Synthesis Markup Language (SSML) Version 1. World Wide Web Consortium, http://www.w3.org/TR/speech-synthesis.

Campbell, N. (2009). An Audio-Visual Approach to Measuring Discourse Synchrony in Multimodal Conversation Data. In *10th Annual Conference of the International Speech Communication Association, Interspeech 2009*, pages 2159–2162, Brighton, United Kingdom.

Cassell, J., Bickmore, T., Billinghurst, M., Campbell, L., Chang, K., Vilhjálmsson, H. H., and Yan, H. (1999). Embodiment in conversational interfaces: Rea. In *Proceedings of the SIGCHI conference on Human factors in computing systems: the CHI is the limit*, pages 520–527, New York, USA.

Cassell, J., Stocky, T., Bickmore, T., Gao, Y., Nakano, Y., Ryokai, K., Vaucelle, C., and Vilhjálmsson, H. (2002). MACK: Media lab Autonomous Conversational Kiosk. In *Proceedings of Imagina '02*.

Clark, H. H. (1996). *Using Language*. Cambridge University Press.

Cohen, W. W. (1995). Fast Effective Rule Induction. In *Proceedings of the Twelfth International Conference on Machine Learning*, pages 115–123. Morgan Kaufman.

De Jong, N. and Wempe, T. (2009). Praat script to detect syllable nuclei and measure speech rate automatically. *Behavior Research Methods*, 41(2):385–390.

De Melo, C. and Gratch, J. (2009). Expression of Emotions Using Wrinkles, Blushing, Sweating and Tears. In Ruttkay, Z., Kipp, M., Nijholt, A., and Vilhjálmsson, H., editors, *Proceedings of the 9th international conference on Intelligent Virtual Agents*, volume 5773 of *Lecture Notes in Computer Science*, pages 188–200. Springer Berlin / Heidelberg.

De Ruiter, J., Mitterer, H., and Enfield, N. (2006). Projecting the end of a speaker's turn: A cognitive cornerstone of conversation. *Language*, 82(3):515–535.

De Sevin, E., Hyniewska, S., and Pelachaud, C. (2010). Influence of Personality Traits on Backchannel Selection. In Allbeck, J., Badler, N., Bickmore, T., Pelachaud, C., and Safonova, A., editors, *Intelligent Virtual Agents*, volume 6356 of *Lecture Notes in Computer Science*, pages 187–193. Springer Berlin / Heidelberg.

De Sevin, E. and Pelachaud, C. (2009). Real-time backchannel selection for ecas according to user's level of interest. In *Proceedings of the 9th International Conference on Intelligent*

Virtual Agents, IVA '09, pages 494–495, Berlin, Heidelberg. Springer-Verlag.

Douglas-Cowie, E., Cowie, R., Cox, C., Amier, N., and Heylen, D. K. J. (2008). The sensitive artificial listener: an induction technique for generating emotionally coloured conversation. In Devillers, L., Martin, J.-C., Cowie, R., Douglas-Cowie, E., and Batliner, A., editors, *LREC Workshop on Corpora for Research on Emotion and Affect, Marrakech, Marokko*, pages 1–4, Paris, France. ELRA.

Duncan, S. and Niederehe, G. (1974). On Signalling That it's Your Turn to Speak. *Journal of Experimental Social Psychology*, 10(3):234–247.

Edlund, J., Heldner, M., and Gustafson, J. (2005). Utterance segmentation and turn-taking in spoken dialogue systems. *Sprachtechnologie, mobile Kommunikation und linguistische Ressourcen*, pages 576–587.

Ekman, P. (1984). Expression and the nature of emotion. *Approaches To Emotion*, pages 1–25.

Ekman, P. and Friesen, W. V. (1978). Facial action coding system: a technique for the measurement of facial movement. *Consulting Psychologists Press*.

Endrass, B., Rehm, M., André, E., and Nakano, Y. I. (2008). Talk is silver, silence is golden: A cross cultural study on the usage of pauses in speech. In *Proceedings of the IUI Workshop on Enculturating Conversational Interfaces (ECI 2008)*.

Eyben, F., Wöllmer, M., and Schuller, B. (2009). openEAR - Introducing the Munich Open-Source Emotion and Affect Recognition Toolkit. In *Proc. 4th International HUMAINE Association Conference on Affective Computing and Intelligent Interaction (ACII 2009)*, pages 576–581, Amsterdam, The Netherlands. IEEE.

Eyben, F., Wöllmer, M., and Schuller, B. (2010). Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the international conference on Multimedia*, MM '10, pages 1459–1462, New York, NY, USA. ACM.

Fontaine, J. R. J., Scherer, K. R., Roesch, E. B., and Ellsworth, P. C. (2007). The World of Emotions is not Two-Dimensional. *Psychological Science*, 18:1050–1057.

Fukayama, A., Ohno, T., Mukawa, N., Sawaki, M., and Hagita, N. (2002). Messages embedded in gaze of interface agentsimpression management with agent's gaze. In *Proceedings of the SIGCHI conference on Human factors in computing systems: Changing our world, changing ourselves*, pages 41–48, Minneapolis, Minnesota, USA. ACM New York, NY, USA.

Giles, H., Taylor, D. M., and Bourhis, R. (1973). Towards a theory of interpersonal accommodation through language: some Canadian data. *Language in Society*, 2:177–192.

Goldberg, J. A. (1990). Interrupting the discourse on interruptions : An analysis in terms of relationally neutral, power- and rapport-oriented acts. *Journal of Pragmatics*, 14(6):883–903.

Goldberg, L. R. (1993). The structure of phenotypic personality traits. *American Psychologist*, 48(1):26–34.

Gratch, J., Okhmatovskaia, A., Lamothe, F., Marsella, S., Morales, M., Van der Werf, R., and Morency, L.-P. (2006). Virtual rapport. In *6th International Conference on Intelligent Virtual Agents*, pages 14–27. Springer.

Gunes, H. and Pantic, M. (2010a). Automatic, dimensional and continuous emotion recognition. *International Journal of Synthetic Emotions*, 1(1):68–99.

Gunes, H. and Pantic, M. (2010b). Dimensional emotion recognition from spontaneous head gestures for interaction with sensitive artificial listeners. In Allbeck, J., Badler, N., Bickmore, T., Pelachaud, C., and Safonova, A., editors, *Proceedings of the 10th International*

Conference on Intelligent Virtual Agents, IVA 2010, volume 6356 of *Lecture Notes in Computer Science*, pages 371–377, Berlin. Springer Verlag.

Hartmann, B., Mancini, M., and Pelachaud, C. (2002). Formational parameters and adaptive prototype instantiation for mpeg-4 compliant gesture synthesis. In *Computer Animation Conference*, pages 111–119.

Jiang, B., Valstar, M. F., and Pantic, M. (2011). Action unit detection using sparse appearance descriptors in space-time video volumes. In *Proceedings of Ninth IEEE International Conference on Automatic Face and Gesture Recognition (FG 2011)*, pages 314–321, Santa Barbara, USA. IEEE.

Johnston, M. (2009). EMMA: Extensible MultiModal Annotation markup language. World Wide Web Consortium, http://www.w3.org/TR/emma.

Jonsdottir, G., Thórisson, K. R., and Nivel, E. (2008). Learning smooth, Human-Like turntaking in realtime dialogue. In *Proceedings of the 8th international conference on Intelligent Virtual Agents*, pages 162–175, Tokyo, Japan. Springer.

Kopp, S., Gesellensetter, L., Krämer, N. C., and Wachsmuth, I. (2005). A Conversational Agent as Museum Guide Design and Evaluation of a Real-World Application. In Panayiotopoulos, T., Gratch, J., Aylett, R., Ballin, D., Olivier, P., and Rist, T., editors, *Intelligent Virtual Agents*, volume 3661 of *Lecture Notes in Computer Science*, pages 329–343. Springer Berlin / Heidelberg.

Kopp, S., Krenn, B., Marsella, S., Marshall, A. N., Pelachaud, C., Pirker, H., Thórisson, K., and Vilhjálmsson, H. (2006). Towards a Common Framework for Multimodal Generation: The Behavior Markup Language. In Gratch, J., Young, M., Aylett, R., Ballin, D., and Olivier, P., editors, *Intelligent Virtual Agents*, volume 4133 of *Lecture Notes in Computer Science*, pages 205–217. Springer.

Landauer, T. K. and Dumais, S. T. (1997). Solution to Plato's Problem: The Latent Semantic Analysis Theory of Acquisition, Induction and Representation of Knowledge. *Psychological Review*, 104:211–240.

Larsson, S. and Traum, D. (2000). Information state and dialogue management in the trindi dialogue move engine toolkit. *Natural Language Engineering*, 6(3–4):323–340.

Leuski, A., Patel, R., Traum, D. R., and Kennedy, B. (2006). Building effective question answering characters. In *Proceedings of the 7th SIGdial Workshop on Discourse and Dialogue*, pages 18–27, Sydney, Australia.

Levin, E., Pieraccini, R., and Eckert, W. (2000). A stochastic model of human-machine interaction for learning dialog strategies. *Speech and Audio Processing, IEEE Transactions on*, 8(1):11–23.

Mancini, M., Bresin, R., and Pelachaud, C. (2007). A virtual-agent head driven by musical performance. *Ieee Transactions On Audio Speech And Language Processing*, 15(6):1833–1841.

Matheson, C., Poesio, M., and Traum, D. R. (2000). Modelling grounding and discourse obligations using update rules. In *Proceedings of the 1st North American chapter of the Association for Computational Linguistics conference*, pages 1–8, Seattle, Washington. Morgan Kaufmann Publishers Inc.

McCauley, L. and D'Mello, S. (2006). MIKI: A Speech Enabled Intelligent Kiosk. In Gratch, J., Young, M., Aylett, R., Ballin, D., and Olivier, P., editors, *Proceedings of the 6th international conference on Intelligent Virtual Agents*, volume 4133 of *Lecture Notes in Computer Science*, pages 132–144. Springer Berlin / Heidelberg.

McKeown, G., Valstar, M., Cowie, R., and Pantic, M. (2010). The Semaine Corpus of Emotionally Coloured Character Interactions. In *Proc. IEEE Int. Conf. on Multimedia & Expo (ICME2010)*, pages 1079–1084, Singapore. IEEE.

Mcrorie, M., Sneddon, I., Sevin, E., Bevacqua, E., and Pelachaud, C. (2009). A model of personality and emotional traits. In *Proceedings of the 9th International Conference on Intelligent Virtual Agents*, IVA '09, pages 27–33, Berlin, Heidelberg. Springer-Verlag.

Mozziconacci, S. (1998). *Speech variability and emotion: production and perception*. Phd thesis, Technical University Eindhoven.

Nijholt, A. and Hulstijn, J. (2000). Multimodal Interactions with Agents in Virtual Worlds. In Kasabov, N., editor, *Future Directions for Intelligent Information Systems and Information Science*, volume 45 of *Studies in Fuzziness and Soft Computing*, pages 148–173. Physica-Verlag, Heidelberg, Germany.

Novick, D., Hansen, B., and Ward, K. (1996). Coordinating turn-taking with gaze. In *Fourth International Conference on Spoken Language Processing*, pages 1888–1891.

O'Connell, D. C., Kowal, S., and Kaltenbacher, E. (1990). Turn-taking: A critical analysis of the research tradition. *Journal of Psycholinguistic Research*, 19(6):345–373.

Op den Akker, R., Bunt, H., Keizer, S., and Van Schooten, B. (2005). From question answering to spoken dialogue - towards an information search assistant for interactive multimodal information extraction. In *9th European Conference on Speech Communication and Technology, Interspeech 2005*, pages 2793–2796, Edinburgh.

Orestrom, B. (1983). *Turn Taking in English Conversation*. Krieger Publishing Company.

Padilha, E. and Carletta, J. (2003). Nonverbal behaviours improving a simulation of small group discussion. In *Proc. 1st Nordic Symp. on Multimodal Comm*, pages 93–105, Copenhagen, Denmark.

Pammi, S. and Schröder, M. (2009). Annotating meaning of listener vocalizations for speech synthesis. *2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops*, pages 1–6.

Pantic, M., Eyben, F., Gunes, H., Heylen, D., Schuller, B., and Wöllmer, M. (2009). Human conversational signals analyser. Technical report, Deliverable D3a, SEMAINE Project.

Quarteroni, S. and Manandhar, S. (2009). Designing an interactive open-domain question answering system. *Natural Language Engineering*, 15(Special Issue 01):73–95.

Raux, A. and Eskenazi, M. (2008). Optimizing Endpointing Thresholds using Dialogue Features in a Spoken Dialogue System. In *Proceedings of SIGdial 2008*, pages 1–10, Columbus, OH, USA.

Rienks, R. J. and Heylen, D. K. J. (2006). Automatic Dominance Detection in Meetings using easily obtainable features. In Bourlard, H. and Renals, S., editors, *Revised Selected Papers of the 2nd Joint Workshop on Multimodal Interaction and Related Machine Learning Algorithms MLMI 2005*, volume 3869 of *Lecture Notes in Computer Science*, pages 76–86, Berlin. Springer Verlag.

Robinson, L. F. and Reis, H. T. (1989). The Effects of Interruption, Gender, and Status on Interpersonal Perceptions. *Journal of Nonverbal Behavior*, 13(3):141–153.

Sacks, H., Schegloff, E., and Jefferson, G. (1974). A simplest systematics for the organization of turn-taking for conversation. *Language*, 50(4):696–735.

Salzberg, S. L. (1994). C4.5: Programs for Machine Learning by J. Ross Quinlan. Morgan Kaufmann Publishers, Inc., 1993. *Machine Learning*, 16(3):235–240.

Sato, R., Higashinaka, R., Tamoto, M., Nakano, M., and Aikawa, K. (2002). Learning decision trees to determine turn-taking by spoken dialogue systems. In *Seventh International Conference on Spoken Language Processing*, pages 861–864, Japan. Citeseer.

Satoh, I. (2008). Context-Aware Agents to Guide Visitors in Museums. In Prendinger, H., Lester, J., and Ishizuka, M., editors, *Proceedings of the 8th international conference on Intelligent Virtual Agents*, volume 5208 of *Lecture Notes in Computer Science*, pages 441–455, Berlin, Heidelberg. Springer Berlin Heidelberg.

Scheffler, K. and Young, S. (2002). Automatic learning of dialogue strategy using dialogue simulation and reinforcement learning. In *Proceedings of the second international conference on Human Language Technology Research*, HLT '02, pages 12–19, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Schegloff, E. (2000). Overlapping talk and the organization of turn-taking for conversation. *Language in Society*, 29(1):1–63.

Schröder, M., Bevacqua, E., Cowie, R., Eyben, F., Gunes, H., Heylen, D., Ter Maat, M., McKeown, G., Pammi, S., Pantic, M., Pelachaud, C., Schuller, B., De Sevin, E., Valstar, M., and Wöllmer, M. (2011). Building Autonomous Sensitive Artificial Listeners (to appear). *IEEE Transactions on Affective Computing*.

Schröder, M., Bevacqua, E., Eyben, F., Gunes, H., Maat, M. T., Pammi, S., Sevin, E. D., Valstar, M., and Wöllmer, M. (2010). Final sal system. Technical report, Deliverable D1d, SEMAINE Project.

Schröder, M. and Cowie, R. (2005). Developing a consistent view on emotion-oriented computing. In *Machine Learning for Multimodal Interaction*, pages 194–205.

Schröder, M., Cowie, R., Heylen, D. K. J., Pantic, M., Pelachaud, C., and Schuller, B. (2008). Towards responsive Sensitive Artificial Listeners. In *Proceedings of the Fourth International Workshop on Human-Computer Conversation*, Sheffield, UK. University of Sheffield.

Schröder, M. and Trouvain, J. (2003). The German Text-to-Speech Synthesis System MARY: A Tool for Research, Development and Teaching. *International Journal of Speech Technology*, 6(4):365–377.

Schuller, B., Vlasenko, B., Eyben, F., Rigoll, G., and Wendemuth, A. (2009). Acoustic Emotion Recognition: A Benchmark Comparison of Performances. In *Proc. Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 552–557, Merano, Italy. IEEE.

Schulman, D. and Bickmore, T. (2009). Persuading users through counseling dialogue with a conversational agent. In *Proceedings of the 4th International Conference on Persuasive Technology Persuasive 09*, Persuasive '09. ACM Press.

Smith, C., Crook, N., Boye, J., Charlton, D., Dobnik, S., Pizzi, D., Cavazza, M., Pulman, S., De la Camara, R., and Turunen, M. (2010). Interaction Strategies for an Affective Conversational Agent. In Allbeck, J., Badler, N., Bickmore, T., Pelachaud, C., and Safonova, A., editors, *Intelligent Virtual Agents*, volume 6356 of *Lecture Notes in Computer Science*, pages 301–314. Springer Berlin / Heidelberg.

Staum Casasanto, L., Jasmin, K., and Casasanto, D. (2010). Virtually accommodating: Speech rate accommodation to a virtual interlocutor. In Ohlsson, S. and Catrambone, R., editors, *Proceedings of the 32nd Annual Conference of the Cognitive Science Society*, pages 127–132, Austin, TX. Cognitive Science Society.

Ten Bosch, L. and Oostdijk, N. (2004). Turn-taking in social talk dialogues: temporal, formal and functional aspects. In *9th International Conference Speech and Computer*, pages 454–461, St. Petersburg.

Ten Bosch, L., Oostdijk, N., and De Ruiter, J. (2004). Durational aspects of turn-taking in spontaneous face-to-face and telephone dialogues. In *Text, Speech and Dialogue*, volume 3206 of *Lecture Notes in Computer Science*, pages 563–570, Brno, Czech Republic. Springer.

Ter Maat, M. and Heylen, D. (2009). Turn management or impression management? In Ruttkay, Z., Kipp, M., Nijholt, A., and Vilhjlmsson, H., editors, *Intelligent Virtual Agents, 9th International Conference, IVA 2009*, volume 5773 of *Lecture Notes in Computer Science*, pages 467–473, Berlin. Springer Verlag.

Ter Maat, M. and Heylen, D. (2010). Generating Simple Conversations. In Esposito, A., Campbell, N., Vogel, C., Hussain, A., and Nijholt, A., editors, *Development of Multimodal Interfaces: Active Listening and Synchrony*, volume 5967 of *Lecture Notes in Computer Science*, pages 92–101. Springer Berlin / Heidelberg.

Ter Maat, M. and Heylen, D. (2011). Flipper: An information state component for spoken dialogue systems. In Vilhjálmsson, H., Kopp, S., Marsella, S., and Vilhjálmsson, K., editors, *Intelligent Virtual Agents*, volume 6895 of *Lecture Notes in Computer Science*, pages 470–472. Springer Berlin / Heidelberg.

Ter Maat, M., Truong, K., and Heylen, D. (2010). How turn-taking strategies influence users impressions of an agent. In Allbeck, J., Badler, N., Bickmore, T., Pelachaud, C., and Safonova, A., editors, *Proceedings of the International Conference on Intelligent Virtual Agents (IVA)*, volume 6356 of *Lecture Notes in Computer Science*, pages 441–453, Berlin. Springer Verlag.

Theune, M., Hofs, D. H. W., and Van Kessel, M. (2007). The Virtual Guide: A direction giving embodied conversational agent. In *Proceedings of the 8th Annual Conference of the International Speech Communication Association (Interspeech 2007)*, pages 2197–2200, Bonn. International Speech Communication Association (ISCA).

Theune, M., Meijs, K., Heylen, D., and Ordelman, R. (2006). Generating expressive speech for storytelling applications. *Audio, Speech, and Language Processing, IEEE Transactions on generating expressive speech for storytelling applications*, 14(4):1137–1144.

Thórisson, K. R. (2002). Natural turn-taking needs no manual: Computational theory and model, from perception to action. *Multimodality in language and speech systems*, pages 173–207.

Traum, D. and Larsson, S. (2003). The Information State Approach to Dialogue Management. In *Current and New Directions in Discourse and Dialogue*, pages 325–353.

Traum, D., Roque, A., Leuski, A., Georgiou, P., Gerten, J., Martinovski, B., Narayanan, S., Robinson, S., and Vaswani, A. (2007). Hassan: A Virtual Human for Tactical Questioning. In *Proceedings of the 8th SIGdial Workshop on Discourse and Dialogue*, pages 71–74.

Traum, D. R., Bos, J., Cooper, R., Larsson, S., Lewin, I., Matheson, C., and Poesio, M. (1999). A model of dialogue moves and information state revision. Technical report, Deliverable D2.1, Trindi-project.

Traum, D. R. and Rickel, J. (2002). Embodied agents for multi-party dialogue in immersive virtual worlds. In *Proceedings of the first international joint conference on Autonomous agents and multiagent systems: part 2*, pages 766–773, New York, USA. ACM.

Trimboli, C. and Walker, M. B. (1984). Switching pauses in cooperative and competitive conversations. *Journal of Experimental Social Psychology*, 20(4):297–311.

Türk, O. and Schröder, M. (2008). A comparison of voice conversion methods for transforming voice quality in emotional speech synthesis. In *Proceedings of Interspeech 2008*, pages 2282–2285. International Speech Communication Association ISCA.

Van Tol, W. and Egges, A. (2009). Real-Time Crying Simulation. In Ruttkay, Z., Kipp, M., Nijholt, A., and Vilhjálmsson, H., editors, *Proceedings of the 9th international conference on Intelligent Virtual Agents*, volume 5773 of *Lecture Notes in Computer Science*, pages 215–228. Springer Berlin / Heidelberg.

Van Zanten, G. V. (1996). Pragmatic interpretation and dialogue management in spoken-language systems. In *Proceedings of the 11th Twente Workshop on Language Technology, TWLT 11*, pages 81–88.

Von der Pütten, A., Krämer, N., and Gratch, J. (2010). How Our Personality Shapes Our Interactions with Virtual Characters - Implications for Research and Development. In Allbeck, J., Badler, N., Bickmore, T., Pelachaud, C., and Safonova, A., editors, *Intelligent Virtual Agents*, volume 6356 of *Lecture Notes in Computer Science*, pages 208–221. Springer Berlin / Heidelberg.

Wai, C., Meng, H. M., and Pieraccini, R. (2001). Scalability and portability of a belief network-based dialog model for different application domains. In *Proceedings of the first international conference on Human language technology research*, HLT '01, pages 1–6, Stroudsburg, PA, USA. Association for Computational Linguistics.

Weizenbaum, J. (1966). ELIZA—a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, 9(1):36–45.

Wollmer, M., Eyben, F., Schuller, B., and Rigoll, G. (2009). Robust vocabulary independent keyword spotting with graphical models. In *Automatic Speech Recognition Understanding, 2009. ASRU 2009. IEEE Workshop on*, pages 349 –353.

Yngve, V. (1970). On getting a word in edgewise. In *sixth regional meeting of the Chicago Linguistic Society*, volume 6, pages 657–677.

Summary

Communication with a machines is inherently not ‘natural’, but still we prefer to interact with them without learning new skills but by using types of communication we already know. Ideally, we want to communicate with a machine just as we communicate with people: we explain (using our voice and gestures) what we want the machine to do, and it understands this and performs the required task.

In its simplest form, such a dialogue system receives the user’s input as written text, which it has to parse and analyze to extract the intentions of the user. But a more complex dialogue system can perceive the user via a microphone and a camera, and the user can use normal speech and gestures to explain his or her intentions. However, this means that the system has to take other aspects of human conversation into account besides interpreting the user’s intentions. For example, it has to manage correct turn-taking behaviour, it has to provide feedback, and it has to manage a correct level of politeness.

This thesis focusses on two aspects of the interaction between a user and a virtual agent (a dialogue system with a visual embodiment), namely the perception of turn-taking strategies and the selection of appropriate responses. This research was carried out in the context of the SEMAINE project, in which a virtual listening agent was built: a virtual agent that tries to keep the user talking for as long as possible. Additionally, the system consists of four specific characters, each with a certain emotional state: a happy, a gloomy, an aggressive, and a pragmatic one. These characters also try to get the user in the same emotional state as they themselves are in.

Turn-taking is a good example of something that is completely natural for most people, but very hard to teach a system. And while most dialogue systems focus on having the agent’s responses start as soon as possible after the user’s end of turn without overlapping it, evidence indicates that starting too early or too late is not always inappropriate per se. People might start speaking too early because of their enthusiasm, or they might start later than usual because they are thinking.

This thesis describes the study of how different turn-taking strategies used by a dialogue system influence the perception that users have of that system. These turn-taking strategies are different start times of the next turn (starting before the user’s turn has finished, directly when it finishes or after a small pause) and different responses when overlapping speech is detected (stop speaking, continue normally or continue with a raised voice).

These strategies were evaluated in two studies. In the first study, a simulator was created that generated conversations by having two agents ‘talk’ to each other. The turn-taking behaviour of each agent was scripted beforehand, and the resulting conversation was played by using non-intelligible speech. After listening to a simulated conversation, the users had to complete a questionnaire containing semantic differential scales about how they perceived a participant in the conversation. In the second study, the users actively participated in the conversation themselves. They were interviewed by a dialogue system, but the exact timing of each question was controlled by a human wizard. This wizard varied the start time of

the questions depending on the selected strategy of that particular interview, and after each interview the users had to complete a questionnaire about how they perceived the dialogue system.

These studies showed that starting too early (that is, interrupting the user) was mostly associated with negative and strong personality attributes: agents were perceived as less agreeable and more assertive. Leaving pauses between turns had contrary associations: it was perceived as more agreeable, less assertive, and created the feeling of having more rapport. It also showed that different strategies influence the response behaviour of the users as well. The users seemed to 'adapt' to the interviewing agent's turn-taking strategy, for example by talking faster and with shorter turns when the interviewer started early during the interview.

The final part of the thesis describes the response selection of the listening agent. We decided to select an appropriate response based on the non-verbal input, rather than on the content of the user's speech, to make the listening agent capable of responding appropriately regardless of the topic. This thesis first describes the handcrafted models and then the more data-driven approach. In this approach, humans annotated videos containing user turns with appropriate possible responses. Classifiers were then used to learn how to respond after a user's turn. Different methods were used to create the training data and evaluate the results. The classifiers were tested by letting them predict appropriate responses for new fragments and let humans rate these responses. We found that some classifiers produced significantly more appropriate responses than a random model.

Samenvatting

Communiceren met machines is per definitie niet ‘natuurlijk’, maar toch zouden we graag met ze communiceren zonder nieuwe vaardigheden te moeten aanleren, maar op een manier die we al kennen. Idealiter zouden we kunnen communiceren met een computer op dezelfde manier als wij communiceren met een ander persoon, namelijk door onze intenties duidelijk te maken door middel van spraak en non-verbaal gedrag.

Bij een simpel dialoogsysteem kunnen gebruikers hun zinnen typen. Deze zinnen worden ontleed en geanalyseerd, en de intenties van de gebruikers worden geëxtraheerd. Een complexer dialoogsysteem kan gebruikers waarnemen met een camera en een microfoon, zodat deze gebruikers gewoon kunnen praten en bewegen om hun intenties over te brengen. Maar dit betekent ook dat het systeem rekening moet houden met andere aspecten van conversaties. Zo zal het bijvoorbeeld moeten zorgen voor nette beurtwisselingen, zal het korte feedback moeten geven terwijl de gebruiker praat, en moet het zorgen voor een bepaalde mate van beleefdheid.

Dit proefschrift focust zich op twee aspecten van interactie tussen een gebruiker en een virtueel persoon (een dialoogsysteem met een visuele belichaming), namelijk de perceptie van verschillende manieren om van beurt te wisselen (dit aspect heet ‘turn-taking’) en het selecteren van passende reacties na de beurt van een gebruiker. Dit onderzoek is uitgevoerd in de context van het SEMAINE project, waarin een virtuele luisteraar is gemaakt. Dit is een virtueel persoon dat als doel heeft om de gebruiker zo lang mogelijk aan de praat te houden. Daarnaast bestaat het systeem ook uit vier specifieke karakters met een eigen emotionele staat: een vrolijk, een depressief, een boos en een pragmatisch karakter. Deze karakters hebben ook het doel om de gebruikers in hun eigen emotionele staat te krijgen.

Turn-taking is een goed voorbeeld van iets dat bij mensen volstrekt natuurlijk gaat, maar moeilijk is om een computer te laten doen. De meeste dialoogsysteem proberen om hun beurt zo snel mogelijk na de beurt van de gebruiker te laten beginnen, maar verschillende onderzoeken wijzen erop dat te vroeg of te laat beginnen van een beurt niet altijd verkeerd is. Soms beginnen mensen te vroeg met praten vanwege hun enthousiasme, or ze beginnen te laat omdat ze nog aan het denken zijn.

Dit proefschrift beschrijft de studies die zijn uitgevoerd over hoe verschillende turn-taking strategieën, gebruikt door een virtueel persoon, de perceptie van een gebruiker verandert. Deze strategieën zijn de verschillende starttijden van een beurt (beginnen voordat de gebruiker klaar is, zo snel mogelijk als de gebruiker klaar is, of na een korte pauze) en het gedrag als zowel de gebruiker en de virtuele persoon tegelijk praten (stoppen met praten, gewoon doorgaan, of doorgaan met een luidere stem).

Deze strategieën zijn geëvalueerd in twee studies. In de eerste studie is een conversatiesimulator gemaakt die een conversatie kan genereren door twee programma’s met elkaar te laten ‘praten’. Het turn-taking gedrag van deze programma’s was vooraf gespecificeerd, en de gegenereerde conversaties werden afgespeeld door middel van onverstaanbare spraak. Nadat gebruikers deze opnames hadden gehoord, moesten ze een vragenlijst invullen met daarin semantische differentiaalschalen over hun perceptie van een deelnemer in de conversatie. In de tweede studie waren de gebruikers actieve deelnemers in het gesprek. Ze werden geïnterviewd door het dialoogsysteem, maar de starttijd van elke vraag werd bepaald door een zogenaamde 'wizard', een persoon op de achtergrond die het dialoogsysteem aanstuurt. De gekozen starttijd varieerde van gesprek tot gesprek, en na elk interview vulde de gebruiker opnieuw een vragenlijst in over hoe de interviewer op hem of haar overkwam.

De uitkomsten van deze studies laten zien dat te vroeg beginnen (dus terwijl de gebruiker nog aan het praten is) werd geassocieerd met negatieve en sterke persoonlijkheidskenmerken: de systemen werden gezien als minder aangenaam en assertiever. Het tegenovergestelde effect werd bereikt door aan het eind van de beurt van de gebruiker pas te beginnen na een korte pauze: het systeem werd dan gezien als aangenamer, minder assertief, en met een grotere betrokkenheid. De resultaten laten ook zien dat verschillende turn-taking strategieën het gedrag van de gebruiker beïnvloedden. De gebruikers pasten zich aan aan de turn-taking strategie van de interviewer, namelijk door sneller te praten met kortere beurten als de interviewer steeds vroeg begon met de vragen.

Het laatste deel van dit proefschrift beschrijft het selecteren van een passende reactie na een beurt van de gebruiker. Om de reacties van de virtuele persoon onafhankelijk te maken van de inhoud van het gesprek is er voor gekozen om de reacties alleen te baseren op het non-verbale gedrag van de gebruiker. Dit proefschrift beschrijft eerst de handgemaakte modellen voor passende reacties, en gaat daarna door met een methode die data gebruikt om zelf goede regels te leren. Voor deze methode hebben een aantal mensen filmpjes met daarin fragmenten van gebruikers bekeken, en voor elk fragment een aantal suggesties voor passende reacties gegeven. Deze suggesties zijn gebruikt als trainingsdata voor een programma dat hiermee zelf regels leert wanneer een reactie passend is en wanneer niet. Deze programma's zijn getest door ze reacties te laten voorspellen voor nieuwe fragmenten en deze te laten beoordelen door mensen. Hier kwam uit dat sommige programma's significant betere reacties produceerden dan een programma dat willekeurig reacties kiest.

SIKS dissertation series

Since 1998, all dissertations written by Ph.D.-students who have conducted their research under auspices of a senior research fellow of the SIKS research school are published in the SIKS Dissertation Series.

2011-47 Azizi Bin Ab Aziz(VU), *Exploring Computational Models for Intelligent Support of Persons with Depression.*
2011-46 Beibei Hu (TUD), *Towards Contextualized Information Delivery: A Rule-based Architecture for the Domain of Mobile Police Work.*
2011-45 Herman Stehouwer (UvT), *Statistical Language Models for Alternative Sequence Selection.*
2011-44 Boris Reuderink (UT), *Robust Brain-Computer Interfaces.*
2011-43 Henk van der Schuur (UU), *Process Improvement through Software Operation Knowledge.*
2011-42 Michal Sindlar (UU), *Explaining Behavior through Mental State Attribution.*
2011-41 Luan Ibraimi (UT), *Cryptographically Enforced Distributed Data Access Control.*
2011-40 Viktor Clerc (VU), *Architectural Knowledge Management in Global Software Development.*
2011-39 Joost Westra (UU), *Organizing Adaptation using Agents in Serious Games.*
2011-38 Nyree Lemmens (UM), *Bee-inspired Distributed Optimization.*
2011-37 Adriana Burlutiu (RUN), *Machine Learning for Pairwise Data, Applications for Preference Learning and Supervised Network Inference.*
2011-36 Erik van der Spek (UU), *Experiments in serious game design: a cognitive approach.*
2011-35 Maaike Harbers (UU), *Explaining Agent Behavior in Virtual Training.*
2011-34 Paolo Turrini (UU), *Strategic Reasoning in Interdependence: Logical and Game-theoretical Investigations.*
2011-33 Tom van der Weide (UU), *Arguing to Motivate Decisions.*
2011-32 Nees-Jan van Eck (EUR), *Methodological Advances in Bibliometric Mapping of Science.*
2011-31 Ludo Waltman (EUR), *Computational and Game-Theoretic Approaches for Modeling Bounded Rationality.*
2011-30 Egon van den Broek (UT), *Affective Signal Processing (ASP): Unraveling the mystery of emotions.*
2011-29 Faisal Kamiran (TUE), *Discrimination-aware Classification.*
2011-28 Rianne Kaptein(UVA), *Effective Focused Retrieval by Exploiting Query Context and Document Structure.*
2011-27 Aniel Bhulai (VU), *Dynamic website optimization through autonomous management of design patterns.*
2011-26 Matthijs Aart Pontier (VU), *Virtual Agents for Human Communication - Emotion Reg-*

ulation and Involvement-Distance Trade-Offs in Embodied Conversational Agents and Robots.
2011-25 Syed Waqar ul Qounain Jaffry (VU)), *Analysis and Validation of Models for Trust Dynamics.*
2011-24 Herwin van Welbergen (UT), *Behavior Generation for Interpersonal Coordination with Virtual Humans On Specifying, Scheduling and Realizing Multimodal Virtual Human Behavior.*
2011-23 Wouter Weerkamp (UVA), *Finding People and their Utterances in Social Media.*
2011-22 Junte Zhang (UVA), *System Evaluation of Archival Description and Access.*
2011-21 Linda Terlouw (TUD), *Modularization and Specification of Service-Oriented Systems.*
2011-20 Qing Gu (VU), *Guiding service-oriented software engineering - A view-based approach.*
2011-19 Ellen Rusman (OU), *The Mind ' s Eye on Personal Profiles.*
2011-18 Mark Ponsen (UM), *Strategic Decision-Making in complex games.*
2011-17 Jiyin He (UVA), *Exploring Topic Structure: Coherence, Diversity and Relatedness.*
2011-16 Maarten Schadd (UM), *Selective Search in Games of Different Complexity.*
2011-15 Marijn Koolen (UvA), *The Meaning of Structure: the Value of Link Evidence for Information Retrieval.*
2011-14 Milan Lovric (EUR), *Behavioral Finance and Agent-Based Artificial Markets.*
2011-13 Xiaoyu Mao (UvT), *Airport under Control. Multiagent Scheduling for Airport Ground Handling.*
2011-12 Carmen Bratosin (TUE), *Grid Architecture for Distributed Process Mining.*
2011-11 Dhaval Vyas (UT), *Designing for Awareness: An Experience-focused HCI Perspective.*
2011-10 Bart Bogaert (UvT), *Cloud Content Contention.*
2011-09 Tim de Jong (OU), *Contextualised Mobile Media for Learning.*
2011-08 Nieske Vergunst (UU), *BDI-based Generation of Robust Task-Oriented Dialogues.*
2011-07 Yujia Cao (UT), *Multimodal Information Presentation for High Load Human Computer Interaction.*
2011-06 Yiwen Wang (TUE), *Semantically-Enhanced Recommendations in Cultural Heritage.*
2011-05 Base van der Raadt (VU), *Enterprise Architecture Coming of Age - Increasing the Performance of an Emerging Discipline..*
2011-04 Hado van Hasselt (UU), *Insights in Reinforcement Learning; Formal analysis and empirical evaluation of temporal-difference.*
2011-03 Jan Martijn van der Werf (TUE), *Compositional Design and Verification of Component-Based Information Systems.*
2011-02 Nick Tinnemeier(UU), *Organizing Agent Organizations. Syntax and Operational Semantics of an Organization-Oriented Programming Language.*
2011-01 Botond Cseke (RUN), *Variational Algorithms for Bayesian Inference in Latent Gaussian Models.*
2010-53 Edgar Meij (UVA), *Combining Concepts and Language Models for Information Access.*
2010-52 Peter-Paul van Maanen (VU), *Adaptive Support for Human-Computer Teams: Exploring the Use of Cognitive Models of Trust and Attention.*
2010-51 Alia Khairia Amin (CWI), *Understanding and supporting information seeking tasks in multiple sources.*
2010-50 Bouke Huurnink (UVA), *Search in Audiovisual Broadcast Archives.*
2010-49 Jahn-Takeshi Saito (UM), *Solving difficult game positions.*
2010-48 Withdrawn, .
2010-47 Chen Li (UT), *Mining Process Model Variants: Challenges, Techniques, Examples.*
2010-46 Vincent Pijpers (VU), *e3alignment: Exploring Inter-Organizational Business-ICT Alignment.*
2010-45 Vasilios Andrikopoulos (UvT), *A theory and model for the evolution of software services.*
2010-44 Pieter Bellekens (TUE), *An Approach towards Context-sensitive and User-adapted Ac-*

cess to Heterogeneous Data Sources, Illustrated in the Television Domain.
2010-43 Peter van Kranenburg (UU), *A Computational Approach to Content-Based Retrieval of Folk Song Melodies.*
2010-42 Sybren de Kinderen (VU), *Needs-driven service bundling in a multi-supplier setting - the computational e3-service approach.*
2010-41 Guillaume Chaslot (UM), *Monte-Carlo Tree Search.*
2010-40 Mark van Assem (VU), *Converting and Integrating Vocabularies for the Semantic Web.*
2010-39 Ghazanfar Farooq Siddiqui (VU), *Integrative modeling of emotions in virtual agents.*
2010-38 Dirk Fahland (TUE), *From Scenarios to components.*
2010-37 Niels Lohmann (TUE), *Correctness of services and their composition.*
2010-36 Jose Janssen (OU), *Paving the Way for Lifelong Learning; Facilitating competence development through a learning path specification.*
2010-35 Dolf Trieschnigg (UT), *Proof of Concept: Concept-based Biomedical Information Retrieval.*
2010-34 Teduh Dirgahayu (UT), *Interaction Design in Service Compositions.*
2010-33 Robin Aly (UT), *Modeling Representation Uncertainty in Concept-Based Multimedia Retrieval.*
2010-32 Marcel Hiel (UvT), *An Adaptive Service Oriented Architecture: Automatically solving Interoperability Problems.*
2010-31 Victor de Boer (UVA), *Ontology Enrichment from Heterogeneous Sources on the Web.*
2010-30 Marieke van Erp (UvT), *Accessing Natural History - Discoveries in data cleaning, structuring, and retrieval.*
2010-29 Stratos Idreos(CWI), *Database Cracking: Towards Auto-tuning Database Kernels.*
2010-28 Arne Koopman (UU), *Characteristic Relational Patterns.*
2010-27 Marten Voulon (UL), *Automatisch contracteren.*
2010-26 Ying Zhang (CWI), *XRPC: Efficient Distributed Query Processing on Heterogeneous XQuery Engines.*
2010-25 Zulfiqar Ali Memon (VU), *Modelling Human-Awareness for Ambient Agents: A Human Mindreading Perspective.*
2010-24 Dmytro Tykhonov, *Designing Generic and Efficient Negotiation Strategies.*
2010-23 Bas Steunebrink (UU), *The Logical Structure of Emotions.*
2010-22 Michiel Hildebrand (CWI), *End-user Support for Access to*
Heterogeneous Linked Data.
2010-21 Harold van Heerde (UT), *Privacy-aware data management by means of data degradation.*
2010-20 Ivo Swartjes (UT), *Whose Story Is It Anyway? How Improv Informs Agency and Authorship of Emergent Narrative.*
2010-19 Henriette Cramer (UvA), *People's Responses to Autonomous and Adaptive Systems.*
2010-18 Charlotte Gerritsen (VU), *Caught in the Act: Investigating Crime by Agent-Based Simulation.*
2010-17 Spyros Kotoulas (VU), *Scalable Discovery of Networked Resources: Algorithms, Infrastructure, Applications.*
2010-16 Sicco Verwer (TUD), *Efficient Identification of Timed Automata, theory and practice.*
2010-15 Lianne Bodenstaff (UT), *Managing Dependency Relations in Inter-Organizational Models.*
2010-14 Sander van Splunter (VU), *Automated Web Service Reconfiguration.*
2010-13 Gianluigi Folino (RUN), *High Performance Data Mining using Bio-inspired techniques.*
2010-12 Susan van den Braak (UU), *Sensemaking software for crime analysis.*
2010-11 Adriaan Ter Mors (TUD), *The world according to MARP: Multi-Agent Route Planning.*
2010-10 Rebecca Ong (UL), *Mobile Communication and Protection of Children.*
2010-09 Hugo Kielman (UL), *A Politiele gegevensverwerking en Privacy, Naar een effectieve*

waarborging.
2010-08 Krzysztof Siewicz (UL), *Towards an Improved Regulatory Framework of Free Software. Protecting user freedoms in a world of software communities and eGovernments.*
2010-07 Wim Fikkert (UT), *Gesture interaction at a Distance.*
2010-06 Sander Bakkes (UvT), *Rapid Adaptation of Video Game AI.*
2010-05 Claudia Hauff (UT), *Predicting the Effectiveness of Queries and Retrieval Systems.*
2010-04 Olga Kulyk (UT), *Do You Know What I Know? Situational Awareness of Co-located Teams in Multidisplay Environments.*
2010-03 Joost Geurts (CWI), *A Document Engineering Model and Processing Framework for Multimedia documents.*
2010-02 Ingo Wassink (UT), *Work flows in Life Science.*
2010-01 Matthijs van Leeuwen (UU), *Patterns that Matter.*
2009-46 Loredana Afanasiev (UvA), *Querying XML: Benchmarks and Recursion.*
2009-45 Jilles Vreeken (UU), *Making Pattern Mining Useful.*
2009-44 Roberto Santana Tapia (UT), *Assessing Business-IT Alignment in Networked Organizations.*
2009-43 Virginia Nunes Leal Franqueira (UT), *Finding Multi-step Attacks in Computer Networks using Heuristic Search and Mobile Ambients.*
2009-42 Toine Bogers (UvT), *Recommender Systems for Social Bookmarking.*
2009-41 Igor Berezhnyy (UvT), *Digital Analysis of Paintings.*
2009-40 Stephan Raaijmakers (UvT), *Multinomial Language Learning: Investigations into the Geometry of Language.*
2009-39 Christian Stahl (TUE, Humboldt-Universitaet zu Berlin), *Service Substitution – A Behavioral Approach Based on Petri Nets.*
2009-38 Riina Vuorikari (OU), *Tags and self-organisation: a metadata ecology for learning resources in a multilingual context.*
2009-37 Hendrik Drachsler (OUN), *Navigation Support for Learners in Informal Learning Networks.*
2009-36 Marco Kalz (OUN), *Placement Support for Learners in Learning Networks.*
2009-35 Wouter Koelewijn (UL), *Privacy en Politiegegevens; Over geautomatiseerde normatieve informatie-uitwisseling.*
2009-34 Inge van de Weerd (UU), *Advancing in Software Product Management: An Incremental Method Engineering Approach.*
2009-33 Khiet Truong (UT), *How Does Real Affect Affect Affect Recognition In Speech?.*
2009-32 Rik Farenhorst (VU) and Remco de Boer (VU), *Architectural Knowledge Management: Supporting Architects and Auditors.*
2009-31 Sofiya Katrenko (UVA), *A Closer Look at Learning Relations from Text.*
2009-30 Marcin Zukowski (CWI), *Balancing vectorized query execution with bandwidth-optimized storage.*
2009-29 Stanislav Pokraev (UT), *Model-Driven Semantic Integration of Service-Oriented Applications.*
2009-28 Sander Evers (UT), *Sensor Data Management with Probabilistic Models.*
2009-27 Christian Glahn (OU), *Contextual Support of social Engagement and Reflection on the Web.*
2009-26 Fernando Koch (UU), *An Agent-Based Model for the Development of Intelligent Mobile Services.*
2009-25 Alex van Ballegooij (CWI), *"RAM: Array Database Management through Relational Mapping".*
2009-24 Annerieke Heuvelink (VUA), *Cognitive Models for Training Simulations.*
2009-23 Peter Hofgesang (VU), *Modelling Web Usage in a Changing Environment.*
2009-22 Pavel Serdyukov (UT), *Search For Expertise: Going beyond direct evidence.*

2009-21 Stijn Vanderlooy (UM), *Ranking and Reliable Classification.*
2009-20 Bob van der Vecht (UU), *Adjustable Autonomy: Controling Influences on Decision Making.*
2009-19 Valentin Robu (CWI), *Modeling Preferences, Strategic Reasoning and Collaboration in Agent-Mediated Electronic Markets.*
2009-18 Fabian Groffen (CWI), *Armada, An Evolving Database System.*
2009-17 Laurens van der Maaten (UvT), *Feature Extraction from Visual Data.*
2009-16 Fritz Reul (UvT), *New Architectures in Computer Chess.*
2009-15 Rinke Hoekstra (UVA), *Ontology Representation - Design Patterns and Ontologies that Make Sense.*
2009-14 Maksym Korotkiy (VU), *From ontology-enabled services to service-enabled ontologies (making ontologies work in e-science with ONTO-SOA).*
2009-13 Steven de Jong (UM), *Fairness in Multi-Agent Systems.*
2009-12 Peter Massuthe (TUE, Humboldt-Universitaet zu Berlin), *Operating Guidelines for Services.*
2009-11 Alexander Boer (UVA), *Legal Theory, Sources of Law & the Semantic Web.*
2009-10 Jan Wielemaker (UVA), *Logic programming for knowledge-intensive interactive applications.*
2009-09 Benjamin Kanagwa (RUN), *Design, Discovery and Construction of Service-oriented Systems.*
2009-08 Volker Nannen (VU), *Evolutionary Agent-Based Policy Analysis in Dynamic Environments.*
2009-07 Ronald Poppe (UT), *Discriminative Vision-Based Recovery and Recognition of Human Motion.*
2009-06 Muhammad Subianto (UU), *Understanding Classification.*
2009-05 Sietse Overbeek (RUN), *Bridging Supply and Demand for Knowledge Intensive Tasks - Based on Knowledge, Cognition, and Quality.*
2009-04 Josephine Nabukenya (RUN), *Improving the Quality of Organisational Policy Making using Collaboration Engineering.*
2009-03 Hans Stol (UvT), *A Framework for Evidence-based Policy Making Using IT.*
2009-02 Willem Robert van Hage (VU), *Evaluating Ontology-Alignment Techniques.*
2009-01 Rasa Jurgelenaite (RUN), *Symmetric Causal Independence Models.*