NICE project (IST-2001-35293)

Natural Interactive Communication for Edutainment

NICE Deliverable D3.7b

Multimodal Output Generation Module for the NICE fairy-tale game

8 December 2004

Authors

Joakim Gustafson and Johan Boye,

Voice Technologies, TeliaSonera

Project ref. no.	IST-2001-35293
Project acronym	NICE
Deliverable status	Public
Contractual date of delivery	1 Nov 2004
Actual date of delivery	8 Dec 2004
Deliverable number	D3.7b
Deliverable title	Multimodal Output Generation Module for the NICE fairy-tale game
Nature	Report
Status & version	Final
Number of pages	21
WP contributing to the deliverable	1
WP / Task responsible	TeliaSonera
Editor	
Author(s)	Joakim Gustafson and Johan Boye
EC Project Officer	Mats Ljungqvist
Keywords	
Abstract (for dissemination)	This report, Deliverable D3.7 from the HLT project Natural Interactive Communication for Edutainment (NICE), describes the construction of the output generation module for the fairy-tale characters. The module receives a richly parameterized semantic instruction from the dialogue manager (WP5) and generates a multimodal output request (verbal utterances, with lip-synchronisation track, additional behavioral instructions to the animated character (face, gaze, body), and manipulations of physical objects in the 3D environment) that are then rendered by the synthesis module and the animation render module.

Table of Contents

1 Introduction

This report describes how the fairy-tale characters' multimodal output is generated in the Nice fairy-tale game. The report is closely related to D4.2, where all possible animations and operations in the Animation Rendition system are described, along with a description of its XML interface. This report refers to Section 2 in Deliverable D1.2-b for the game scenario and its implications for multimodal output. The report is also naturally coupled to deliverable D5.2-b, which describes the dialogue manager for the fairy-tale characters.

2 Background

Humans who engage in face-to-face dialogues use non-verbal communication such as body gestures, gaze, facial expressions and lip movements to transmit information, attitudes and emotions. If computers are to engage in spoken dialogue with humans it would seem natural to give them the possibility to use non-verbal communication too. An embodied conversational character could increase the believability of the system and make the interaction more natural. Studies have shown that users who interact with an animated talking agent spend more time with the system, enjoy the interaction more and think that the system performed better. This has been called the *persona effect*, and it is considered by many researchers to be the most important reason for adding animated agents in educational systems (Walker et al. 1994, Koda and Maes 1996, Lester et al. 97, van Mulken 1998, Lester et al. 1999). Reeves and Nass (1996) have shown that users tend to interact socially with computers in the same way as they interact with people even though the system does not have a human appearance. Laurel (1990) and Cassell et al. (1999) argue that interface designers could take advantage of anthropomorphism by embodying some types of interfaces, thus making the interaction more natural. Finally, it has been found that embodied conversational characters can make the dialogue situation more entertaining and engaging, (Dehn and van Mulken 2000).

2.1 Conversational skills

It is important for embodied conversational characters to have conversational skills. They have to be able to communicate their goals and plans to the user, and they should be able to cooperate with the user to solve problems. In order to convey personality and to build a collaborative trusting relationship with the users, the characters also have to be able to engage in socializing *small talk* (Cassell & Bickmore, 2002). In order to be able to coordinate their action towards a goal that is shared with the user, the characters have to be able to *collaborate* with the user (Grosz, B. and Sidner, 1986). The characters also have to be able to engage in *grounding* dialogue with the users to be able to certify that they have understood what the user wants them to do. Allwood et al (1991) describe four basic communicative functions that correspond to the levels suggested by Clark (1994) at which problems can arise during the grounding process:

level 1 – contact, vocalization and attention
level 2 – perception, identification
level 3 – understanding, meaning
level 4 – attitudinal reaction , proposal and uptake

In conversation the coordination of turns is crucial. Allwood (1995) defines a turn as a speaker's right to the floor, and suggests that this right is regulated by a number of turn management sub-functions that can be expressed verbally or non-verbally (i.e. eye gaze and head nods). There are two simultaneous information channels in a dialogue: the information channel from the speaker, and the back-channel feedback from the listener. The back-channel feedback indicates attention, feelings and understanding, and its purpose is to support the interaction (Yngve, 1970). It is communicated by anything from short vocalizations like "*mm*" to utterances like "*I think I understand*", or by facial expressions and gestures (Goodwin, 1981). There are a number of studies of linguistic adaptation during error resolution, both on the acoustic level (increases in duration and pitch range, hyperarticulation, louder and often with clearer articulation) and on higher linguistic levels (rephrase of a previous turn, use of words like "*no*" and explicit clarification meta utterances like "*what do you mean?*"). Brennan and Hulteen (1995) have suggested a feedback model for human-computer interaction that builds on a grounding process. A conversational system should provide positive feedback in successful contexts and negative feedback when problems have been detected. Byron and Heeman (1997) describe how the initial cue words (discourse markers) "*and*", "*oh*", "*so*" and "*well*" are used to facilitate grounding by providing information on the speaker's orientation towards the content of the previous turn. Disfluencies like filled pauses may be indicators of problems in dialogue, but they have also been found to be useful for the listener in processing the speaker's turns. Initial fillers are used to manage turn taking, pauses and fillers indicate feeling-of-knowing, and some fillers like "*uh*" can speed monitoring of a subsequent word (Brennan, 2000).

2.2 Non-verbal capabilities

Animating the face brings the embodied character to life, making it more believable as a dialogue partner. According to Ekman (1979) facial actions can be clustered according to their communicative functions in three different channels: the phonemic, the intonational and the emotional. *The phonemic channel* is used to communicate redundant and complementary information in what is being said. Fisher (1968) coined the term *viseme* for the visual realization of phonemes. Accurate lip movements in audiovisual speech can improve intelligibility (Benoît et al. 1994). *The intonational channel* is used to facilitate a smooth interaction. Facial expressions, eyebrow raising and head nods can be used to communicate the information structure of an utterance, for instance stressing new or important objects (Scherer 1980, Pelachaud et al. 1994, Cassel et al. 2001, Decarlo et al. 2002). *The emotional channel* is used to increase the animated character's social believability. Ekman et al. (1972) found the six universal emotions. There are *display rules* that regulate when speakers show emotions. These rules depend on the meaning the speaker wants to convey, the mood of the speaker, the relationship between speaker and listener and the dialogue situation (Ekman 1982).

According to Kahneman (1973) gaze indicates three types of mental processes: spontaneous looking, task-relevant looking and looking as a function of orientation of thought. Thus, in conversation gaze carries information about what the interlocutors are focusing on. Gaze can be used to communicate the speaker's degree of attention and interest during a conversation, to regulate the turn-taking, to refer to visible objects, to show the speaker's mental activity, to display emotions or to define power and status. Pelachaud et al. (1996) described a facial animation system that among other things could display different gaze patterns. According to Duncan (1972) speakers can give cues that indicate the end of their turns not only with prosody and syntax, but also by changing the direction of their gaze. According to Goodwin (1981) the listener looks away from the speaker while taking the turn to avoid cognitive overload while planning what to say. The usefulness of gaze in turn-handling was investigated by Cassell et al. (1999). They found that the speakers looked away from the listeners at the beginning of turns and towards the listeners at the end of turns. They also found that speakers tended to look away from the listeners while giving old information (theme) and towards the listeners while giving new information (rheme). If theme coincided with the start of a turn, the speakers always looked away from the listeners. Thórisson (2002) describes a turn-taking model called the Ymir Turn-Taking Model (YTTM) that uses speech detection, prosody, gesture and body language to determine when the animated agent should take the turn. The BEAT system uses gaze, head nods and eyebrow-raising for turn-handling (Cassel et al. 2001). Finally, according to Colburn et al. (2000) turn-handling gaze can be used to indicate who is talking in multi-party dialogues such as virtual conferencing.

Rimé and Schiaratura (1991) have presented a classification system with six classes of gesture usage in dialogues. *Speech markers (beats, batons)* are used to communicate the information structure of an utterances, e.g. to stress important or new objects in a verbal utterance. *Ideographs* are produced while the speaker is preparing an utterance to indicate the direction of thought. *Iconic gestures* are used to show some representation of an object that is being referred to verbally. The gesture can depict the shape, some spatial relation or action of an object. *Pantomimic gestures* play the role of the referent. *Deictic gestures* are used to point to objects visual in the users' environment or represented in the graphical interface. Finally, *Emblematic gestures* are gestures that have a direct translation into words that is known in a specific culture or social group. They are used to send messages like thumbs up for "ok".

3 The different kinds of multimodal character outputs

The fairy-tale characters in the NICE system are able to generate both verbal and non-verbal behaviour. The non-verbal behaviour include: *physical action*; *emotional display*; *state of mind, turn regulation cues, back-channeling gestures* and *different kinds of body movements*. The characters' verbal behaviour include*: plot dependent speech acts, social exchanges* and *general dialogue regulating speech acts*. The characters in the fairy-tale world have different roles in the game and consequently they have to be able to convey different personalities that match their respective roles. Charles and Cavazza (2004) distinguish between two types of characters in their character-based story telling system - *feature characters* and *supporting characters*. In the Nice fairy-tale game a third kind of character have been added - a *helper character*. This means that there are these three types of characters in the fairy-tale world:

Helper character — A character that guides and helps the user throughout the whole fairy-tale game. In the second prototype *Cloddy Hans* is the Helper. He is a friendly character with no long-term goals for himself, other than doing what the user asks him to.

Feature characters — Characters that has a key function in the plots. In the second prototype *Karen* is a feature character that has a *Gatekeeper* function in the second scene. She is a selfish character with goals of her own. She will not help the user unless she gets a reward.

Supporting characters — Characters that only tell the pieces of information needed for the plot, but that are not willing to engage in conversation with the user. In some cases they function as an obstacle in the plot. In the second prototype Thumbelina will have a minor obstacle role.

These different types of characters need different levels of output capabilities. *Helper characters* need conversational output capabilities allowing both for grounding and cooperation, and for dialogue regulation and error handling. Furthermore, they need to be able to perform general body gestures and actions, as well as special gestures and physical actions that are needed to help the user to solve tasks in the scenes. *Feature characters* need less cooperative and grounding verbal abilities, since they have goals of their own that they simply want to convey to the user. However, they need dialogue regulation and error handling output capabilities. Furthermore, they only need the body gestures and actions that are needed for their functions in their scenes. S*upporting characters* only need to be provided with the output capabilities needed to convey the information they are supposed to communicate to the user. Apart from these they only need to be provided with verbal utterances like "I don't want to chat with you" "Hurry up and do something". They will therefore not be provided with the conversational skills that the other two types of characters will have. In the few cases where the supporting characters are supposed to perform specific physical actions, these will be implemented as cut-scene animations. All three kinds of characters are provided with idle behaviours used when not interacting with the user.

An important role of the fairy-tale characters' multimodal output is to convey their different personalities. Personality is conveyed by modes of appearance, voice quality, choice of words and actions. There are a number of psychological models of personality, one of the most used is the OCEAN model that has the following dimensions: *Openness, Conscientiousness*, *Extroversion, Agreeableness,* and *Neuroticism* (McCrae & Costa 1996). Personality traits are not explicitly modelled in the NICE system, but rather used as guidance in the design of the characters. This means that there is not a set of rules that determines the characters' output behaviour from personality parameters. Instead the personality descriptions are used as a tool to get consistent character behaviours that are perceived by the users as compatible with the intended personality of the characters.

The two main characters in the second prototype of the fairy-tale game are the helper Cloddy Hans and the gatekeeper Karen. In order to match their different roles in the game, the output behaviour of these characters have been designed to display these quite different personality traits:

	Openness	**Conscientiousness**	**Extroversion**	**Agreeableness**	**Neuroticism**
Cloddy Hans	Dunce	Thorough	Quiet Uncertain	Friendly Polite	Calm Even-tempered
Karen	Intellectual	Frivolous	Outspoken Self-confident	Unfriendly Touchy	Anxious Over-emotional

Figure 1. The OCEAN personality traits of the two main characters in the second scene.

The two characters' dialogue rules, wording of utterances, speaking styles and non-verbal behaviour have been designed to match their respective personalities.

4 The verbal output capabilities of the characters

The fairy-tale characters are able to talk about the plots and scenes, as well as their own plans and to goals that relate to these. When characters first meet the user they are able to engage in formalized socializing small talk. In this social phase the characters also have the conversational goals to tell the user some pieces of information about themselves. In later phases they are still able to respond to social initiatives from the users, but without goals of their own to pursue the social topic. Finally, the characters are provided with general dialogue regulating speech acts that they can use in all scenes. An overview of different types of dialogue acts and the included speech acts is given in Figure 2.

Plan Regulating	Error Handling	Turn Handling	Attitudinal Feedback	Discourse markers	Extralinguistic sounds
agree | disagree	report (hearing | not hearing)	feedback continuer	(positive | negative) (filler words | filled pause)	restate	clear throat
(ask for | accept| reject | correct) request	report (understanding | not understanding | misunderstanding)	floorholder (easy | hard) question	yes | no	summarize plan	cough
(accept | reject) offer	report (knowing | not knowing)	backchannel question	apology | non-apology	request for summary	exhalation
	report (correct | wrong) action	neutral filled pauses	grateful | ungrateful	conclude	inhalation
	error acknowledgement		attitude to (grateful | ungrateful)	elaborate plan	laughter
	ask for (clarification | repetition | rephrase)		attitude to (success | failure)	correction	sigh
	(open | bounce) question		attitude to (good | bad)	respond to unexpected info	

Figure 2. An overview of general dialogue regulating speech acts that will be used in all scenes.

The plan regulating acts include utterances like *"What do you want me to do?"*, the error handling acts include *"Could you repeat that?"*, turn handling utterances include *"okay"* and filled pauses, the attitudinal feedback includes both phrases like *"too bad"* and filled pauses with encouraging prosody, the discourse markers include phrases like "*like I said"* and cue words like "*oh"*. Apart from these speech acts, there are also a number of extralinguistic sounds like "*laughter"* and "*sigh"* that can be used to indicate the characters' state of mind or current attitude. In order to be able to talk about the plot, their goals and plans, the fairy-tale characters have been provided with a number of task oriented plot dependent speech acts:

- Introduction and explanation of the plot
- Initiatives that serve to fulfill the characters' plan or long-term goals.
- Requests for new instructions. (e.g."*where do you want me to go*" "so, *what do you want me to do with this axe?*")
- Responses to instructions from the user.
    - Confirmations (e.g. "*Do you want me to pick up the axe?").*
    - Accepts (e.g. "*Ok, I will pick up the axe!*").
    - Rejection (e.g. "*I don't want to lower the drawbridge*"
    - Clarification questions following incomplete interpretations (e.g. "*Where did you want me to place the axe?*" or "*I did not understand which object you meant, could you point at it?*")
    - Clarification questions following conflicts on what goal to pursue (e.g. User: "*Pick up the knife!*" Cloddy "*But, you told me that you wanted me to pick up the AXE?")*
- Stating intentions, plans and goals
    - Upcoming actions (e.g. "*I am going to the shelf now*")
    - Plans to fulfill goals (e.g. "*I will lower the drawbridge if I get something for it*")
    - Long term goal (e.g. "*I want to get over to the other side*")
- Answer user questions
    - Social (e.g. "*I am 30 years old*")
    - Info-questions (e.g. "*It is the farmer's house*")
    - Explaining actions (e.g. "I'*m going to the shelf in order to be able to pick up the axe*")

5 The non-verbal behaviour of the characters

The main fairy-tale characters are able to engage in conversation with the users and perform the actions needed to progress through the plots of the game. In order to be believable as life-like characters they are able to react to user input as well as to events in their environment. These reactions are either displays of attitude (emotional displays, see Ekman et. al. 1972), state of mind (listening, thinking and understanding), turn regulation gestures (turn taking or back-channelling gestures) or attention gestures which can be used when the user start to speak or gesture as well as if another character starts to speak or if something happens in the environment. The characters can also look at and point at interactive objects (moveable objects, draw bridge), non-interactive objects (nearby houses, trees and roads) and landmarks (forests and hills) in the 3D-world. They are able to walk between locations that are far apart. The characters do not stand completely still at one location forever – if the user has not engaged in interaction with them for a while they enter an idle state where they start off with small encouraging gestures, then after a while they indicate impatience by gazing around in the environment or displaying various idle gestures. Finally if the user seems to be unwilling to communicate with them they start walking about in the scene, performing different kinds of non-communicative gestures and actions. The pictures below shows the different types of non-verbal behaviour the characters are able to display.

Physical actions

goTo(drawbridge) **turnTo(book)** **pickUp(sword)**

Emotional display

neutral **surprise** **anger** **happiness** **sadness**

State-of-mind gestures

idle **listening** **thinking** **not understanding** **not hearing**

Back-channelling gestures

raising eyebrows   small nodding

Turn regulation feedback gestures

attention (while away walking)   attention (while standing looking at user)   continued attention   taking turn

Specific body movements

falling   whispering   crossing arms   gripping

General movements of a single body part

close eyelids   turn torso right   turn head right   lift right arm side

The gestures, movements and actions of the different characters are also used to convey their respective personalities. The intended personality traits of the two main characters were given in section 2, and the corresponding OCEAN parameters, they can be summarized as:

Cloddy Hans – low Openness, high Conscientiousness, low Extroversion, high Agreeableness, low Neuroticism
Karen – high Openness, low Conscientiousness, high Extroversion, low Agreeableness, high Neuroticism

To make the characters' output behaviour consistent, the body gestures, actions and idle behaviours of the two characters have been designed with their respective personality traits in mind. The manner in which characters move conveys their different personalities in the same way as their different speaking styles does. The manner of movements can be described using Laban Movement Analysis (LMA, Laban 1971) in terms of Shape and Effort, where Effort is comprised of four components:

Space – attention to the surroundings
Weight – sense of the impact of one's movement
Time – lack or sense of urgency
Flow - toward bodily tension and control.

Chi et al (2000) has developed a parametrisized system, EMOTE, that is based on the LMA method. Allbeck and Badler (2004) describes an initial attempt to link the EMOTE parameters with the OCEAN personality parameters. If this linkage is applied to the two characters, the following EMOTE parameters are obtained for them:

Cloddy Hans - Space(direct), Weight(strong), Time(sudden), Flow(bound)
Karen - Space (indirect), Weight (light), Time(sustained), Flow (free)

The effect of these EMOTE parameters on the non-verbal behaviour of the two main characters is summarized in Figure 3.

	Space	Weight	Time	Flow
Cloddy Hans	*Direct:* Single focus, e.g. he either looks bluntly at the user, or glances at the object that he or the user is referring to.	*Strong*: Powerful, having impact, e.g. he walks with determined steps towards his goal.	*Sudden*: Hurried, urgent, e.g. he performs the actions the user want him to do immediately	*Bound:* Controlled, restrained, e.g. he walks the shortest way to a location, and then he turns to the user, looking encouraging.
Karen	*Indirect*: Multi-focus, e.g. she doesn't look at the user for a very long time, before breaking their mutual gaze, letting her gaze wonder into the surroundings.	*Light*: Delicate, easily overcoming gravity, e.g. she walks about with light steps.	*Sustained*: Lingering, indulging in time, e.g. she tries to avoid to do what the users asks her as long as possible	*Free*: Uncontrolled movement, e.g. she wanders about on her way to an location, looking as she doesn't quite know where she is heading

Figure 3. The impact of the derived EMOTE parameters on the characters' non-verbal behaviours.

The characters' non-verbal behaviours are controlled by the *Animation Handler* module. It sends requests to the *Animation Renderer*, telling it either to play animations and/or sounds or to perform certain character actions. The professional animators at Liquid Media have provided all characters with a number of communicative gestures, as well as a number of simple, single body part animations that can be used to generate more complex multi body part gestures. This makes it possible to either play ready animations for communicative gestures, like the ones for "Thinking", "Did not understand" and "Did not hear" that Karen shows in Figure 4, or to generate animation lists consisting animation tracks on the individual body parts, like the one Cloddy Hans shows for the same set of gestures.

Thinking **Did not understand** **Did not hear**

Figure 4. Two personality-dependent sets of state -of-mind gestures.

To support the intended personalities of these characters, the introvert, shy, quiet, calm, polite Cloddy Hans displays small, but slow and deliberate body gestures that are generated by the AnimationHandler, while the outspoken, self-confident, touchy and anxious Karen displays larger, and faster body gestures, as rendered by professional animators. The characters' different personalities are also conveyed by their different idle behaviours: Karin is less patient than Cloddy Hans, which is reflected by the fact that she enters the idle phase faster and has more complex idle gestures. Cloddy Hans keeps his attention at the user, while Karin's attention from time to time wanders away from the user to the environment. If the user has not said anything for a long time Karen eventually walks away from the user and she starts strolling around aimlessly in the back of the scene, looking at trees and flowers. When she has walked away from the drawbridge, the user has to talk to her to get her attention again, in order to get her to return and continue the conversation. Since both characters have been provided with the same sets of animation it is possible to switch their non-verbal behaviours. This makes it possible to perform user studies on the effect of (in)consistency between the verbal and non-verbal behaviours in terms of the personality they convey. Cloddy's smaller idle gestures are mainly located in the face, while Karen moves her whole body, see Figure 5.

Figure 5. Two personality-dependent sets of idle behaviours.

6 Non-character actions and operations

The system can influence objects in the 3D world by means of the following operations:

- render/not render object
- set position of object
- highlight object
- object-specific actions (highlight/move one of the slots in machine, play error-sound while opening rejection compartment)
- change viewpoints also referred to as cameras with different positions and directions
    - set the active camera (switch to the camera instantly)
    - interpolate to a camera (fly to the camera smothly)
    - set target entity (set the object that a follow-camera should follow)

These operations are used to initialise scenes in the game, placing objects and characters in their initial positions. The users interact with Cloddy Hans to manipulate objects and to move between locations. Highlighting of objects is used to feedback the result of the interpretation of the users' gestural inputs, as well as for emphasizing deictic utterances by Cloddy Hans. The camera angle switches in cases where it is necessary to zoom in on Cloddy Hans or certain objects in order do see them clearly. The camera position changes to make it possible to follow Cloddy Hans as he walks from the shelf with objects to the fairy-tale machine. For a complete list of operations possible to request from the Animation Renderer see deliverable D4.2-2.

7 Generation of character animations and actions

The character Animation System is able to generate all types of verbal and non-verbal behaviours that are described above. Furthermore, the fairy-tale characters need to be both autonomous and reactive at the same time. This means that the characters have to be able to generate carefully planned goal-oriented actions as well as very fast, less planned actions (e.g. attention indicating gestures). In order to be able to build a system that can harness all these functionalities, an event driven, asynchronous, modular system architecture was chosen. An overview of all components in this system architecture for the fairy-tale game system can be found in D3.6. The *Animation System* in Figure 5 is responsible for generating the character animations and actions. It is divided into two modules: The *Animation Handler* and the *Animation Renderer:*

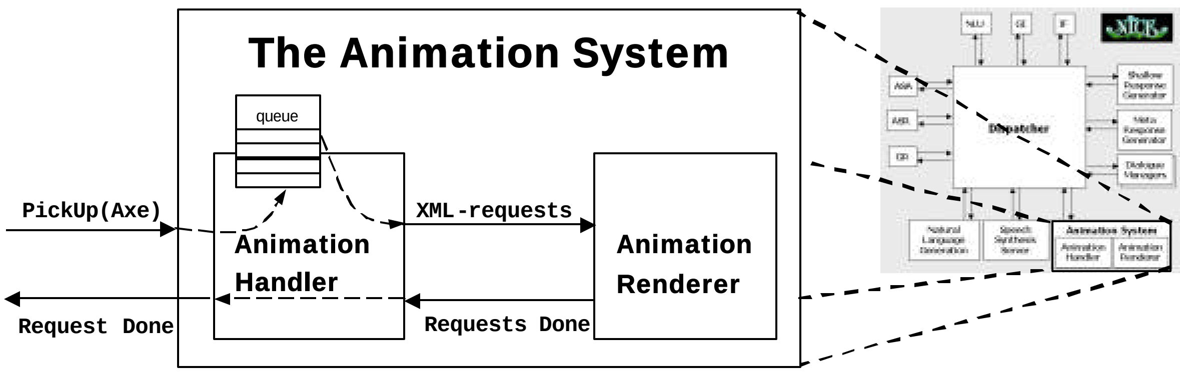


Figure 6. The internal handling of requests to the Animation System and its place in the system.

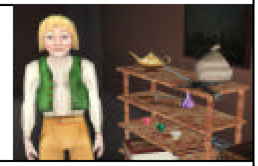
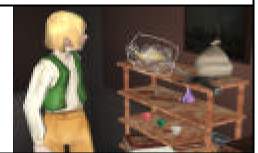
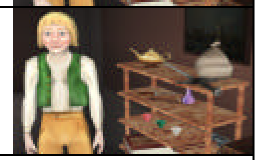
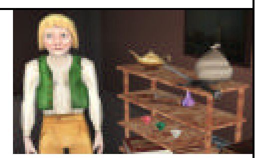
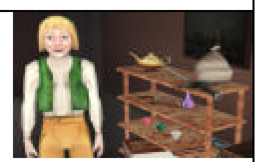
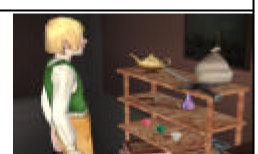
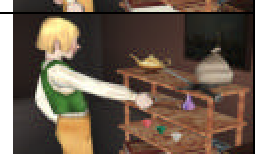
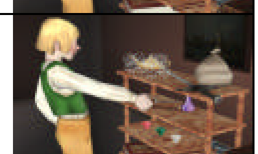
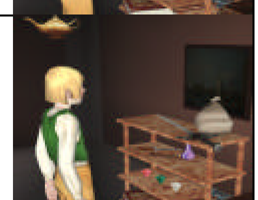
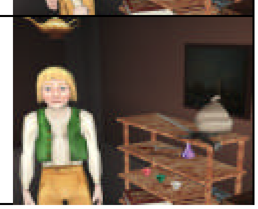
The Animation Handler translates action requests from the Dialogue Manager into animation and action requests that it sends to the Animation Renderer for execution. The Animation Handler has an internal animation queue and can construct complex actions that involve more than one request to the Animation Renderer. The Animation Handler tells the synthesizer to generate a sound file with corresponding lip-synchronization track, and to timestamp animation tags that have been inserted into the text string. It then generates animation requests that are coordinated with the verbal part, and sends the multi-track animation request to the Animation Renderer for execution.

The Animation Renderer executes the animations and actions that the Animation Handler requests and it informs the Animation Handler when its requests have been performed. It also generates trigger events when walking characters enters trigger areas, see D4.2 for details. The interface between the Animation Handler and the Animation Renderer is XML-coded, and a description of the format of that interface along with the available animations and actions can be found in D4.2. When the Animation System has finished an animation or action it informs the Message Dispatcher, which in turn informs the Dialogue Manager that its earlier request has been performed.

The chapter about the fairy-tale game system in D3.6 listed all messages that the Messages Dispatcher routed between the system modules. In order to illustrate the event driven character behaviour we will, starting from a particular user input, trace all messages that are generated from the various system modules, and list the actions that the Animation System performs as a result of some of these. Let us assume that the user inputs the following multimodal message:

<An encircling gesture around the Axe> <0.5 seconds of silence> “Pick this up!”

This will lead to a number of messages between the system modules. Triggered by some of these, the Animation Handler will generate character actions and behaviours.

System Module(Message)	Animation Handler Request	Animation Renderer Request
GR(StartOfGesture)	AttentionTo(GestureInput)	1) RaiseEyebrows 
GR(EndOfGesture)		
GI(Select(Lamp))	LookAt(Lamp)	2) Play(Animation(TorsoTurnLeft(amount), HeadTurnLeft(amount))) 
ASR(StartOfSpeech)	AttentionTo(SpeechInput)	3) TurnTo(ActiveCamera) 
ASR(EndOfSpeech)		
ASA(EndOfTurn)	TakeTurnGesture	4) Play(Animation(Eyegaze(UpRight), Eyebrow(frown))) 
NLP(PickUp(Lamp)) (by simple fusion)		
DM(convey(tell(PickUp(Lamp))))		
NLG(convey("I will pick up the lamp")	Say("I will pick up the lamp") (the AnimationHandler first requests the Synthesizer to generate the sound file and lip-synch track)	5) Play(Sound(synthesis_file) Animation(Lipsynch_track)) 
Dispatcher(2 seconds timeout)		
DM(perform(PickUp(Lamp)))	PickUp(Lamp)	6a) TurnTo(Lamp) 
		6b) Play(Animation(PickupGesture)) 
		6c) Highlight(Lamp) 
		6d) PickUp(Lamp) 
		6e) TurnTo(ActiveCamera) 

The Animation Handler has to keep track of the locations of the cameras, the characters and the objects in order to generate animation requests like 2) above, turning a character's head and torso towards an object. It also has to keep track of which actions the characters are currently performing in order to generate different kinds of attention behaviour requests like the ones in 1) and 3). Figure 6 below, shows two different types of attention behaviours used in two different situations.

Attention (while away walking) **Attention (while standing looking at user)**

Figure 6. Two attention behaviours that depend on the characters current state.

The request from the DM to perform the complex action of picking up the lamp is translated into five simpler actions (6a-e), that the Animation Handler puts in its internal animation queue. The Animation Handler first sends an animation/action request to the Animation Renderer. It then waits for a message from the Animation Renderer that it has finished executing the request before it takes the next action from the queue and sends that. The Animation Handler is also able to queue up incoming animation requests from the other modules. The *convey* (say utterance) request have a special queuing functionality, where different kinds of utterances have different priorities. It is also possible to add an append tag to the convey request. This allows users to barge-in and interrupt the characters, but also makes it possible for the characters to say a number of utterances after each other. The following table shows the priorities and append tags used for utterance requests from different sources:

source of the synthesis request	example utterance	priority	append
user initiated barge-in(no!) from the meta response generator	Sorry, I misunderstood	30	no
task-oriented convey from DM	Do you want me to go to the axe?	30	yes
social utterance from DM	my name is Cloddy	20	no
social utterance from shallow response generator	thanks for being so nice	10	no
encouraging utterances generated from DM as result of a no speech timeout	Maybe we should start building a story soon.	0	no

The combined effect of the priorities and append-tags are as follows: Let us assume that the system is currently saying utterance A, when a request to say utterance B comes in to the Animation Handler. If B has higher priority it will always interrupt A and say B immediately. If B has lower priority than A, B will never be said at all. If A and B has the same priority the existence of an append-tag will decide the functionality. If there is an append-tag in the B request, B will be said as soon as the character has finished saying A. If there is no append-tag in the B request, then A will be interrupted, and B will be said immediately. This makes it possible to generate a burst of task-oriented utterances in a turn, but it also makes it possible for the user to barge-in and cancel a long utterance that was generated due to a misunderstanding. Furthermore, it prevents the system to queue up a large number of responses to user input that were generated by the shallow response generator.

To give the characters basic simple perceptual abilities a number of reactive behaviours have also been added in the Animation System:

Auditory perception is simulated by generating attention gestures that for example involve turning to the speaker. When the ASR has generated a StartOfSpeech event the characters will turn to the active camera, and when there are multiple character speaking in a scene the other characters will turn towards the speaking character (i.e. when the AnimationRenderer starts playing a synthesised character utterance).

Visual perception is simulated by generating attention gestures when the users starts gesturing or glancing at the object that the user has encircled. It is also simulated by adding triggers nearby interesting objects, and letting the Animation Handler generate an appropriate attention gesture towards an object that the character walks by. Another way to simulate simple visual perception is the ability to request a list of all objects that are visible (either on the screen or from a characters field of vision), and then request the character to turn to a found object. Both these kinds of visual perception cues will also be routed to the Dialogue Manager, making it possible for it to plan actions and verbal contributions that relate to nearby objects.

Perception of time is simulated by letting the Dispatcher generate timeouts that inform the Animation Handler that a certain amount of time has passed since the last user input or system output. The Animation Handler keeps track of all character and object locations, as well as the characters' current actions, in order to be able to change a certain character's behaviour dependent on the current situation or the actions it is currently performing, and to be able to coordinate different characters' simultaneous actions.

The Animation Renderer will generate trigger events as soon as a character walks into a trigger area. This feature is used in the fairy-tale world to generate walk paths between locations that are far apart. Between the locations a number of triggers are located at places where the character is to change his direction. When the Animation Handler wants a character to walk from location A to location B, it sends a request to the Animation Renderer to make the character go to the first trigger on the walk path between the A and B. As soon as the character enters the first trigger area the Animation Renderer sends a *characterEnteredTriggerArea1* message to the Animation Handler. It immediately sends back a request to the Animation Renderer that the character should go to the second trigger instead. This is done before the character has reached its destination (the centre of the first trigger area), which means that the character continues walking, but changes it direction towards to the second trigger. This will be repeated until the character reaches location B, as a result of which the character walks on a smooth, non-straight path from A to B.

8 Task oriented natural language generation

The basic stages of utterance generation are as follows:

1. Deciding the content of the utterance, expressed as one or several domain-dependent dialogue acts. (Compare Deliverable D1.2b, and Deliverable D5.2b.) Note: We assume that all dialogue acts have a verbal realization.
2. Determining a) which output modalities to be used and b) the kinds of anaphoric expressions and deictic gestures to be used, if any.
3. Surface realization of the verbal utterance as text.
4. Speech synthesis of the verbal utterance.
5. Rendering of deictic gestures, if present.

Stages 1–2 and 3 can be taken to correspond to the standard distinction in text generation between text planning and linguistic realization, respectively. Since stages 1–2 require access to the dialogue context, they are carried out by the dialogue manager (see Deliverable D5.2b). Stage 3 is carried out by the surface-generation module. Stages 4 and 5 are carried out by the speech-synthesis and animation modules, respectively. The latter three stages are described below.

A more fine-grained division would also include "aggregation" of utterances, that is, deciding how to fold several elements of information into one or more sentences, (Reiter and Dale 1997). This is also relevant here, but for the time being we simply concatenate the realizations of dialogue acts.

The input to stage 3 (surface generation) is one or more domain-dependent dialogue acts, using the same representation as that used for final analysis of utterances by the dialogue manager, except that substructures not to be realized can be hidden (see further below).

The dialogue-act representation carries implicit information about the "verbosity" with respect to (pieces of) the utterance, that is, the degree to which linguistic material is included or left out (as in elliptic utterances). This allows us to produce different realizations from the same underlying representation depending on what is appropriate in the given context. Lack of verboseness is coded by "folding" (hiding) substructures not to be realized in the output; the corresponding substructures are enclosed in a term whose functor is a "+" sign. Below we give an example of realizations of an unfolded dialogue-act representation, and a number of ways of folding parts of the content.

```
tell(cloddy, user, intend(cloddy, putdown(cloddy, axe, shelf)))
```

"I'm going to put the axe on the shelf"

```
tell(cloddy, user, intend(cloddy, putdown(cloddy, +(axe), useful)))
```

"I'm going to put it on the shelf"

```
tell(cloddy, user, intend(cloddy, putdown(cloddy, axe, +(useful))))
```

"I'm going to put the axe there"

```
tell(cloddy, user, intend(cloddy, +(putdown(cloddy, axe, useful))))
```

"I'm going to do it"

The actual generation is implemented by means of a Definite Clause Grammar (see Sterling and Shapiro 1994). The dialogue act considered previously:

```
tell(cloddy, user, intend(cloddy, putdown(cloddy, axe, shelf)))
```

is turned into its surface realization "Jag ska lägga yxan på hyllan" (I'm going to put the axe on the shelf") by the following rules:

```
generate_dialogue_act( tell(cloddy, user, C) ) -->
  generate_proposition( C ).

generate_proposition( intend(cloddy, putDown(cloddy, A, B)) ) -->
  [jag,ska,lägga],
  generate_object( A ),
  generate_location( B ).

generate_object( axe )      --> [  yxan ].

generate_location( shelf )  --> [ på, hyllan].
```

All in all, the generation grammar contains some 200 rules.

9 Synthesizing the verbal output

In commercial spoken dialogue systems a lot of effort is put into persona design of the voices. These systems mostly use recorded prompts to obtain naturalness and quality. In the Nice, system the helper character can engage in collaborative task-oriented dialogues that include clarification system utterances. This makes the number of possible system utterances too large to be pre-recorded. In limited domains it is possible to build natural sounding synthetic voices using a unit selection synthesizer, like Festival/FestVox (Black 1998, Black & Lenzo 2000b). Limited domain synthesizers (Black & Lenzo 2000a) have been developed for a number of applications, e.g., in a task oriented dialogue system in the travel domain (Rudnicky et al 2000) and for animated characters in a military training domain (Johnson et al 2002). The limited domain synthesis approach was decided to be the most promising method for creating voices for the fairy-tale characters in the NICE project. In order to be able to build Swedish voices with natural voice quality and prosody, a unit selection synthesizer was developed in cooperation with KTH(Gustafson&Sjölander 2004).

The Synthesizer server receives requests from the AnimationPlanner which contain the verbal utterances generated by the NLG server, with inserted animation tags that contains aniamation ID used by the AnimationPlanner to generate the non-verbal behaviour. The Synthesizer produces a sound file containing the verbal realization of the utterance along with a lip-synchronization animation track. It also time stamps the animation tags in the animation track and sends them back to the AnimationHandler that the generate the timestamped animations in synchrony with the verbal output that the synthesizer generated.

To ensure a fast and responsive system, cached synthesized utterances (sound files together with lip-synchronisation tracks) are used during runtime. Different methods are used to generate the cached utterances for the fairy-tale characters. The first method is to generate utterances in batch mode, using a text file with all possible character outputs as input. The second method to generate cached utterances by running the system in wizard mode, where it is possible for the wizard to type in utterances that have not been prepared for in advance. In this case, all utterances that get synthesized by the unit selection synthesizer are automatically cached and stored in files. The task-oriented natural language generator is capable of generating a large number of utterances (about 6000). These involve all possible actions that can be performed with the fairy-tale objects and all possible actions that Cloddy Hans can perform.

It would be very inefficient to store 6000 full utterances when the utterances have a clear slot-filling structure. Instead a prompt building functionality has been included in the Synthesis module. The reason for this is both that is it faster than the unit selection synthesizer and that it makes it possible to determine exactly which prompt fragment to use in a certain case. This could either be a fragment with a desired continuation prosody, or a fragment that ends with a filled pause (indicating uncertainty). In order to handle the task oriented utterances, 300 carrier sentences were designed with the following structure:

1. okey <SILENT PAUSE> I will put <FILLED PAUSE> <SILENT PAUSE> the key in <SILENT PAUSE> dangerous
2. because <SILENT PAUSE> I will put <SILENT PAUSE> the key in <FILLED PAUSE> <SILENT PAUSE> magical
3. do you want me to go to <FILLED PAUSE> <SILENT PAUSE> > the shelf
4. so <FILLED PAUSE> <SILENT PAUSE> then I will go to <SILENT PAUSE> the shelf <SILENT PAUSE> again

This gave a number of prompt parts that are final or not final, where non-final parts ends with a continuation prosody or with a filled pause with continuation prosody, like these:

1: **I_will_put** (not final, ending with a filled pause with continuation prosody)
2: **I_will_put** (not final, ending with continuation prosody)
3: **the_shelf** (final)
4: **the_shelf** (not final, ending with continuation prosody)

By the time of writing the generation module has hand-coded tags for *emphasis, cue words* and *pauses* in certain generation rules. This will be replaced by a mechanism where the dialogue state and confidence scores from the understanding modules will be used to determine when and where these tags will be inserted.

An important role of the synthesis component in the fairy-tale system is to convey the personality of the characters. The two main characters have the personality traits:

Cloddy Hans – low Openness, high Conscientiousness, low Extroversion, high Agreeableness, low Neuroticism
Karen – high Openness, low Conscientiousness, high Extroversion, low Agreeableness, high Neuroticism

Nass and Lee (2000) showed that it was possible to design a synthetic voice with stereotypical extrovert vocal features (high loudness, increased pitch, a large frequency range and a fast speaking rate) that in fact was described by most users with extrovert personality adjective items. Cloddy Hans and Karen have been provided with voices with speaking styles that are designed to be as suitable as possible for their respective personality traits:

	Voice pitch	Speaking rate	Frequency range
Cloddy Hans	Low	Slow	Small
Karen	High	Fast	Large

Figure 7. The overall speaking styles of the two main characters.

To get to the different speaking styles, the voice talents were told to read the utterances in manners that matched the targeted personalities. This resulted in two voices with speaking styles that, among other things, differed in frequency range. They also differed in speaking rate and voice pitch. In order to accentuate these last two differences, all utterances were re-sampled changing speaking rate and voice pitch at the same time. All Cloddy's utterances were slowed down and all Karen's utterances were speeded up. This simple procedure had desired side-effects: apart from making Cloddy's voice slower it made him sound larger, and, apart from making Karen's voice faster, it made her sound younger. The personalities of the two characters were deliberately chosen so that this simple voice transformation would also make their voices more matching with the visual appearance of the two animated characters.

Future work includes adding mechanisms to change the realisation of the verbal and non-verbal behaviour depending on the dialogue state, as well as depending on the confidence scores of the understanding modules' interpretation of the user's previous turn. Furthermore we will try to verify that the characters' intended personalities are conveyed to the users in user studies. In the user studies performed so far 41 users have only interacted with Cloddy Hans, and all of them have indeed described him as slow, stupid, friendly and helpful. However, we anticipate that it will be harder to convey Karen's fast, smart and unfriendly personality, which is what we intend to study in the next set of user studies. We will also investigate how the characters' personalities along with their limited understanding capabilities influence the users' appreciation of the fairy-tale game as a whole. Finally, we will study the effects of inserting filled and unfilled pauses as well as discourse markers in the verbal output of Cloddy Hans.

10 References

Allbeck, J. and Badler, N. "Representing and Parameterizing Agent Behaviours" In Life-Like Characters: Tools, Affective Functions, and Applications, Prendinger, H. and Ishizuka, M, Eds. Springer, Germany, 2004.

Allwood, J., Nivre, J. and Ahlsén, E. "On the Semantics and Pragmatics of Linguistic Feedback," in Journal of Semantics, 1992.

Allwood, J. "Reasons for management in dialog", in Beun, R.J., Baker, M. and Reiner, M. (eds.) Dialogue and Instruction. Springer-Verlag.pp 241-50, 1995.

Arafa et al (2002) "Two approaches to Scripting Character Animation", http://www.vhml.org/workshops/AAMAS/papers/Kamyab.pdf

Benoît, C. & Sawallis, T. R. (eds.), *Talking machines: Theories, Models, and Designs*, North Holland, Elsevier, Amsterdam, pp. 505–522.

Beskow J, Edlund J, Nordstrand M "A model for generalised multi-modal dialogue system output applied to an animated talking head", in Minker, W., Bühler, D. and Dybkjaer, L. (eds) Spoken Multimodal Human-Computer Dialogue in Mobile Environments, Dordrecht, The Netherlands, Kluwer Academic Publishers, 2003

Black A., Taylor P., and Caley R., "The Festival speech synthesis system," http://festvox.org/festival. 1998.

Black, A. and Lenzo, K. "Limited Domain Synthesis", ICSLP2000, Beijing, China, 2000a.

Black, A. and Lenzo, K. "Building Voices in the Festival Speech Synthesis System," http://festvox.org/bsv, 2000b.

Brennan, S. "Processes that shape conversation and their implications for computational linguistics," Proceedings, 38th Annual Meeting of the ACL. Hong Kong, 2000.

Brennan, S. and Hulteen, E. "Interaction and feedback in a spoken language system: a theoretical framework," Knowledge-Based Systems(8): 143-151, 1995

Byron, D. and Heeman, P. "Discourse Marker Use in Spoken Dialog", In Proceedings of the 5th Eurospean Conference On Speech Communication and Technology, Rhodes, Greece, September 1997.

Cassell, J., Vilhjálmsson, H. & Bickmore, T. "BEAT: The Behavior Expression Animation Toolkit". Proceedings of SIGGRAPH '01, Los Angeles, CA, 2001, pp. 477–486, 2001

Cassell, J., Bickmore, T. "Negotiated Collusion: Modeling Social Language and its Relationship Effects in Intelligent Agents,", User Modeling and Adaptive Interfaces(12) 1-44, 2002.

Cassell, J., Bickmore, T., Billinghurst, M., Campbell, L., Chang, K., Vilhjlmsson, H. & Yan. H. "Embodiment in conversational interfaces: Rea". Proceedings of the Association for Computing Machinery (ACM) Special Interest Group on Computer–Human Interaction (SIGCHI), Pittsburgh, PA., May 1999, pp. 520–527.

Charles, F. and Cavazza, M. "Exploring the Scalability of Character-based Storytelling". ACM Joint Conference on Autonomous Agents and Multi-Agent Systems, New York, USA, 2004

Chi, D., Costa, M., Zhao, L. and Badler, N. "The EMOTE Model for Effort and Shape," Proceedings of SIGGRAPH 2000, ACM Computer Graphics Annual Conference, New Orleans, Louisiana, 23-28 July, 2000, pp. 173-182.

Clark, H. "Managing problems in speaking", Speech Communication, 15:243-250, 1994.

Colburn, A., Cohen, M. & Drucker, S. (2000) "The Role of Eye Gaze in Avatar Mediated Conversational Interfaces". MSR-TR-2000-81. Microsoft Research, 2000.

Dehn, D. & van Mulken, S. "The impact of animated interface agents: a review of empirical research". International Journal of Human– Computer Studies, vol. 52, pp. 1–22.

Duncan Jr., S. “Some signals and rules for taking speaking turns in conversations”. Journal of Personality and Social Psychology, 23(2), pp. 283– 292, 1972.

Ekman, P. *Emotion in the human face*. New York: Cambridge University Press, 1982.

Ekman, P. “About brows: Emotional and conversational signals”. In: Cranach, M. von, Foppa, K., Lepenies, W. & Ploog, D. (eds.), Human Ethology: Claims and Limits of a New Discipline: Contributions to the Colloquium, Cambridge, Cambridge University Press, pp. 169–202, 1979

Ekman, P., Friesen, W.V., Ellsworth, P. *Emotion in the Human Face: Guidelines for Research and an Integration of Findings*. New York, Pergamon Press Inc, 1972.

Fisher, C. “Confusions among visually perceived consonants”. Journal of Speech and Hearing Research 11, pp. 796–804, 1968.

Goodwin, C. “Conversational Organization: interaction between speakers and hearers,” New York/London, Academic Press, 1981.

Grosz, B. and Sidner, C. “Attention, Intention, and the Structure of Discourse,” Computational Linguistics 12(3), pp. 175–204, 1986.

Johnson, W., Narayanan, S., Whitney, R. Das, R., Bulut, M. and LaBore, C. “Limited Domain Synthesis of Expressive Military Speech for Animated Characters,” In Proceedings of the IEEE TTS Workshop, 2002.

Kahneman, D. *Attention and Effort*. Prentice–Hall, Englewood–Cliffs, New Jersey, 1973.

Koda, T. & Maes, P. “Agents with faces: The effects of personification of agents” Proceedings of Human–Computer Interaction, London, pp. 239–245, 1996.

Laurel, B. “Interface agents : Metaphors with character”. In: Laurel, B. (ed.), The Art of Human– Computer Interface Design, Addison–Wesley, pp. 355–365, 1990.

Lester, J. Voerman, J. Towns, S. & Callaway, C. “Deictic Believability: Coordinated Gesture, Locomotion, and Speech in Lifelike Pedagogical Agents”. Applied Artificial Intelligence, 13(4–5), pp. 383–414, 1999.

Lester, J., Converse, S., Kahler, S., Barlow, S., Stone, B. & Bhogal, R. “The persona effect: Affective impact of animated pedagogical agents”. Proceedings on Human Factors in Computing Systems, Atlanta, GA, pp. 359–366, 1997.

McCrae, R. and Costa, P. “Toward a new generation of personality theories: Theoretical contexts for the five-factor model,” In J. S. Wiggins (Ed.), The five-factor model of personality: Theoretical perspectives (pp. 51-87). New York: Guilford, 1996.

Nass, C. and Lee, K. “Does computer-generated speech manifest personality? An experimental test of similarity-attraction,” Proceedings of the CHI 2000 Conference, The Hague, The Netherlands, 2000.

Pelachaud, C. & Prevost, S. “Sight and Sound: Generating Facial Expressions and Spoken Intonation from Context”. Proceedings of the second ESCA Workshop on Speech Synthesis, New Paltz, NY, USA, September 1994, pp. 216–219, 1994.

Pelachaud, C., Badler, N. & Steedman, M. “Generating Facial Expressions for Speech”. Cognitive Science 20(1), pp. 1–46, 1996.

Perlin, K. and Goldberg, A. “Improv: A system for scripting interactive actors in virtual worlds. Proceedings of SIGGRAPH ’96 pp205-216, 1996.

Reeves, B. & Nass, C. *The Media Equation: how people treat computers, televisions and new media like real people and places*, Cambridge, Cambridge University Press, 1996.

Reiter, E. and Dale, R. “Building applied natural language generation systems”, Natural Language Engineering, 3:57--87, 1997. http://citeseer.nj.nec.com/reiter97building.html

Rimé, B. and Schiaratura, L. “Gesture and speech”. In: Feldman, Robert S. / Rimé, Bernard (eds): Fundamentals of nonverbal behavior. Cambridge : Cambridge Univ. Pr. pp. 239-281, 1991.

Rudnicky, A., Bennett, C., Black, A., Chotomongcol, A., Lenzo, K., Oh, A., Singh, R. “Task and domain specific modelling in the Carnegie Mellon Communicator system,” Proceedings of ICSLP, 2000.

Scherer, K. “The functions of nonverbal signs in conversation”. In: Giles, H. & St. Clair, R. (eds.), The Social and Physiological Contexts of Language, 1980.

Sterling L. and Shapiro E. *The art of Prolog*, 2nd edition. The MIT Press, 1994.

Van Mulken, S, André, E. and Müller, J. ”The Persona Effect: How Substantial Is It”, In Proceedings of HCI'98, pages 53-66, Sheffield, UK, 1998.

Thórisson, K. “Natural Turn-Taking Needs No Manual: Computational Theory and Model, from Perception to Action”. In: Granström, B., House, D., Karlsson, I. (eds.), *Multimodality in Language and Speech Systems*, Dordrecht, The Netherlands, Kluwer Academic Publishers, pp. 173–207, 2002.

Walker, J., Sproull, L. & Subramani, R. “Using a human face in an interface”. Proceedings of the CHI ‘94 Conference Companion on Human Factors in Computing Systems, Boston, MA, pp. 85–91, 1994.

Yngve, V. ”On getting a word in edgewise,” In Papers, from the Sixth Regional Meeting, Chicago Linguistic Society, pages 567–577, 1970.