Requirements for Spoken Conversational Agents in Games

Term Paper in the GSLT-1 course Speech Technology, autumn 2005

Jenny Brusk
jenny.brusk@hgo.se

30th January 2006

Abstract

Spoken conversational systems have mainly been designed to collaborate with a user to solve a particular task, such as managing bank transactions or book travels. Now the development has reached a point where new areas of application may be possible, such as in digital games and interactive story worlds. The aim of this paper is to present a number of features that are required if we are to place a spoken conversational agent (SCA) in a virtual game world. We will explore the differences between a *typical* SCA and an *in-game* SCA, and see what consequences these difference have on the requirements.

1 Introduction

Research on spoken conversational systems has up to now mainly focussed on task-oriented systems and other useful applications that aim to help people and companies in their daily life. Now the development has reached a point where new areas of application may be possible, such as in digital games and interactive story worlds. There are several research projects aiming at integrating natural language processing with virtual worlds; Façade, an interactive story which allows the player to conversate with the two characters in the drama using typed text (Mateas and Stern, 2004), The NICE Fairy-tale Game Scenario (Gustafson et al., 2005), in which conversational characters can communicate verbally and non-verbally with the player about things and events in the game world, and Character-based Interactive Story by Cavazza and Charles (2005) and Mead et al. (2003), where the user is regarded as an active spectator with the ability to change the main character's plan.

The idea of introducing natural language dialogues in virtual game worlds leaves us with the following questions:

- In what type of games and game contexts do we find it relevant to be able to converse with a character in natural language?

- What can we talk about? What is relevant?
- What player types can we identify based on how they communicate with the system?
- Should the conversation have in-game consequences, such as changing the relationship between participants in the conversation or changing the state of affairs in the game (in Façade for instance, the player's interaction may change the dramatic beat (Mateas and Stern, 2002) and thereby the story, and in The Sims™, conversations affect the relationship between the participant in the dialogue)
- What is needed in order to create a believable character in a virtual game world with the ability to talk using natural language?

The topic of this paper is to investigate the requirements that need to be fulfilled in order to place a spoken conversational agent (SCA) in a game world, which means that we will focus on the last question throughout the paper.

1.1 Spoken Conversational Agents and Dialogue Systems

Spoken conversational agents are programs that can communicate with humans in natural language (Jurafsky and Martin, 2000). They have mainly been designed to collaborate with the user to solve a particular task, such as managing bank transactions or book travels, i.e. capable of handling dialogues which Allen et al. (2001) describe as *practical*, for example SJ's voice-based travel service. These dialogues are usually limited to a specific domain, which has the advantage that the range of possible interpretations is limited which therefore increase the robustness of these systems. Recent research show more complex examples, such as embodied conversational agents with the ability to small talk and gesture as well as collaborate to solve a task (for instance Waxholm (http://www.speech.kth.se/waxholm/waxholm2.html), August (http://www.speech.kth.se/august/) and Adapt (http://www.speech.kth.se/ctt/proj/adapt/), all developed at the department of Speech, Hearing and Music at KTH, Stockholm, and The NICE project, in which the deceased author H.C. Andersen has come alive in a virtual remake of his study, and may have multi-modal communication with users about things in the spatial context, his authorship, his life but can also small talk to some extent (such as asking the user about his/her age).

The research within the field of conversational agents is now moving towards integrating emotions, personality, cognition, perception and other human-like qualities in the aim to develop a virtual human (Egges et al., 2003; Gratch et al., 2002; Swartout et al., 2004).

Gustafson (2002) suggests classifying spoken dialogue systems in three categories based on the type of dialogue they can handle rather than the technology used (such as the classifications proposed by for instance Allen et al. (2001) and McTear (2002)): *task-oriented dialogues*, which have a limited domain and

usually last for a short time, *explorative dialogues*, in which the overall goal has been removed and in which the agent instead have the ability to help the user with goals that are more difficult to define, and finally, *context-oriented dialogues*, which allows the user to ask and talk about the agent's personality, the spatial context and the situation.

We will henceforth refer to *typical* SCAs, by which we will mean SCAs that are capable of handling task-oriented dialogues.

1.2 Dialogues and Communication in Games

This study will concentrate on games placed in virtual worlds, such as digital games and hybrid or trans-reality games (games that are partly physical, partly virtual, see Lindley (2004) for a more thorough investigation), but first we will give a short description of how dialogues and other communicative actions usually work in games.

Most digital games, like movies and literature, are designed using a combination of recurrent patterns (see Björk and Holopainen (2005) for reference) and where each combination may define the genre of the game. Some genres, such as roleplaying games and adventure games, let the player or player character "talk" to the non-player characters in the game and this interaction is an important part of progressing the game. The conversations are scripted and part of the story of the game. In order to control the flow of the game, the player may choose what to say from a menu and most utterances have some function apart from just being social or informative. For instance, they may serve the purpose of trading or delivering quests.

In games played by thousands of simultaneous players, so called massively multiplayer online games (MMOGs), players play against and *with* each other. These games are highly social and also designed to support socialization (Ducheneaut and Moore, 2004), both in the choice of game features, such as professions, meeting places (e.g. cantinas), quests and guilds, as well as in the more obvious communication channels; chat, dialogues and emotes (menu-based gestures and utterances, see example figure 1). However, these system are mainly text-based, even if the non-player characters may use (recorded) voice in the interaction. Chat and emotes do not in itself halt the player's current activities in the game, but since the player needs to activate them using the mouse or the keyboard, other actions may have to be postponed. In a situation of high collaboration with other characters in the game, for instance a group of characters collaborating in killing a monster, this of course is a disadvantage.

Today, voice-control utilities are available for games, such as Game Commander®, Game Voice and VR Commander. Apart from handling commands specified and defined by the user (basically corresponding to keystrokes, see for instance Game Commander manual http://www.gamecommander.com/misc/gc2.pdf), they can also be used for voice-based chat. These systems thus serve the purpose of allowing the user to do several simultanous tasks, such as using a combination of keystrokes, speech and mouse control.

Figure 1: Emote menu from Disney's multiplayer online game Toontown (children's game)

Figure 2: Dialogue menu from Morrowind (roleplaying adventure game)

Following from the above, speech in game environments can serve different purposes:

Voice-based chat, metacommands and other interface commands. Today already served by voice-control systems. They do not correspond to any dialogue situation and therefore not important for the present paper.

Functional dialogue. Dialogues between a player character and non-player character that serves a particular purpose in the game, such as trade, deliver quests or guide the player. These dialogues are controlled by the system, scripted and leaves only a number of options for the player to choose from (see for example figure 2). If we were to classify these dialogues according to Gustafson (2002), we can see that they have elements of all three categories, but since we have decided to define them as *functional*, the task-oriented element as well as the explorative element will cover most of the interaction.

In-game conversations By which we mean conversations between the player or player character and non-player characters in the game. The main characteristic of these dialogues is that they are social rather than functional. The participants may small talk, discuss personal matters, such as relations, personality and emotions, as well as talk about things in the shared spatial context. They may or may not have a specific purpose or defined goal, it depends on the purpose of initiating the dialogue, and the purpose or goal need not be a specific task to solve. These dialogues could be classified as context-oriented dialogues according to Gustafsons classification. In The Sims^{TM}2, for example, the characters can simulate talk, displayed as icons wrapped in a speech bubble above the character's head. These icons symbolise different topics, or rather *dialogue acts* that are either generated by the system or chosen by the player. The player can however only choose dialogue act for a marked (chosen) character that s/he controls. The dialogue acts can be verbal, for instance *say goodbye* as well as non-verbal, such as *flirt, kiss* and *hit* (see for example 3). The dialogues will affect the relationship between the participants and may have a strong impact on the events and actions that follow (a dialogue may for instance have the consequence that a couple get divorced and separate).

The aim of this paper is to compare the requirements needed to develop a *typical* SCA as opposed to an *in-game* SCA (according to the above classification).

2 Requirements for Spoken Conversational Agents

Spoken conversational agents are used for different purposes and depending on the requirements for the system, some features may be more important than others. In this section we will present different features for spoken conversational agent.

Figure 3: Dialogue menu from The Sims^{TM}2 (simulation game)

2.1 Speech and Language Technology

For spoken conversational agents, speech technology is of course required. Speech technology includes the automatic processing of speech input (speech recognition and/or understanding) as well output (speech synthesis).

Variability Speech recognition for in-game conversations, can be problematic for several reasons; speaker variation, environmental (and even perhaps in-game) sound disturbances and variation in the quality of the input channel. Most of these problems are general for spoken conversational agents as well, but there are other problems such as the fact that most games are not localised, but released in English only. The language skill of the players can differ significantly; there may be variations in glossary, pronunciation and grammatical skills which can make speech recognition hard. We can also expect the the environmental conditions to vary, some play at internet cafés, others play at home and the surrounding noise may range from almost nothing to loud and intensed. The input channel may also vary, it is possible to use the built-in microphone, but also headsets with varying microphone quality. It can therefore be expected to be difficult to set the conditions for training the ASR system.

Incremental speech processing Humans process speech input (an output) incrementally, which allows us to for instance barge-in and give feedback during the conversation. The introduction of incremental text/speech processing also

raise the issue of turn-taking, when is it OK to take the turn? There are also other dialogue features that can only be modelled using incremental speech processing, for instance errors made by the speaker. In human conversations, it is common for the speaker to make errors, such as false-starts, hesitations and ungrammatical constructions. If the aim is to simulate realistic human conversation, including the errors made, rather than building a task-oriented system, incremental speech processing is required.

Robustness and Error Handling A speech recognizer tries to analyze what the speaker said using either a statistical language model or a grammar. If a grammar is used, the recognizer need well-formed grammatical utterances from the user in order to understand what the user is saying. One problem that comes with this approach is that users rarely produce grammatically well-formed sentences, instead they contain errors represented by for instance false-starts, self-repair, hesitations and ungrammatical sentences. One way to handle this problem is to use partial parsing, i.e. to identify well-formed chunks rather than whole sentences (McTear, 2002).

A language model (LM) use statistics retreived from training a large speech corpus. One such technique is the N-gram model that predict the next word in the sequence, based on the previous $N-1$ words. The advantage of using statistical LM is that it is based on how people actually talk and the larger the corpora, the better the analysis.

There are still different problematic events to handle such as *no-input*, when the user is silent instead of answering the system's request or question, and *no-match*, when the user says something that the system cannot understand. Skantze (2003) differentiate between mis-understandings and non-understandings when humans fail in their communication. A mis-understanding occurs when one of the participants makes an interpretion that doesn't match the speaker's intention, whereas a non-understanding refers to situations when the addressee fails to obtain any interpretation at all. The system must be able to decide whether to do an interpretation of a noisy input or treat it as a non-understanding. According to Skantze (2003), the system must at least return a partial result with confidence score, from which the system can decide whether to do an interpretation or treat the input as a non-understanding.

There are different strategies for handling speech recognition errors: in W3C recommendations (http://www.w3.org/TR/voicexml20/) errors such as **noinput** and **nomatch** are handled as *catch elements*, where the default action is to reprompt. A more sophisticated solution is to ground (give feedback and acknowledge) the speaker's utterance when the system decides upon interpretation rather than non-understanding (Skantze, 2003). Do we need additional or other strategies for handling speech recognition errors in games? Should we perhaps individualize the strategies for the characters, making them part of the personality? E.g. letting some characters ignore utterances that they don't fully understand, while other characters are more cooperative and use different grounding methods in order to understand and give feedback. These questions

need to be studied further.

2.2 Interaction and Dialogue Features

In this section we will describe typical (human) dialogue features, and how they are dealt with in dialogue systems and conversational agents.

Initiative In human dialogues, any of the participants may initiate and/or control the dialogue. Dialogue systems are either system-initiative, i.e. the system initiates and controls the dialogue, user-initiative, where the user controls the interaction (in systems that use graphical user interfaces, the user is commonly in charge), or mixed-initiative, meaning both system and user may lead the interaction and the initiative may switch during the interaction. An example of a mixed initiative system is MIMIC, "a voice-enabled telephone-based dialogue system..., that adapts dialogue strategies based on participant roles, characteristics of the current utterance, and dialogue history" (Chu-Carroll, 2000). MIMIC uses *adaptive* mixed initiative, which means that the system automatically adapts the distribution of initiative based on the information extracted from user utterances and dialogue history. The conclusion to draw from this study is that mixed-initiative require more parameters to consider than just being able to take initiative, for instance the participants' interpersonal roles as well as the dialogue history, task and goal.

Intuitively, a in-game SCA must have mixed-initiative, which also is confirmed by Barbara Hayes-Roth in (Hirsh, 1998), where she gives seven principles for character-based interactive stories, amongst which mixed initiative forms one such principle. The participants must be able to decide when and with whom to interact.

Multi-party Dialogue Humans do not only speak with one person at a time, at occasion there are reasons to adress several persons at a time, such as asking a company on the street for a direction or to adress one person but with other potential hearers. Traum (2004) discusses several issues to consider when going from a two-party dialogue to a multi-party dialogue, for instance who is adressed and who can recieve an utterance? It can also be difficult to identify the speaker, for instance in situations when the communicators cannot see each other. Turn-taking (see more below) is also problematic, since there are more agents competing for the turn. Typical SCAs do not have a need for multi-party dialogue, one user typically interact with one agent at a time in managing the task. In a virtual world, however, either habitated by virtual humans (see for example Swartout et al. (2004)) or game characters, multi-party dialogues become an important feature in the attempts to simulate real interaction. In The Sims^{TM}, for instance, the characters can spontaneously engage in a multi-party dialogue, it cannot be initiated by the player. The different dialogue options in The Sims, resulting in a simulated conversation, corresponds to certain *acts* that have impact on the game progression, and they are an important factor in the story that emerges within the game.

Turn-taking In dialogues there are at least two participants engaged in a conversation. However, the participants rarely speak at the same time (*overlap*), rather, they seem to know exactely when to take the turn. If several participants are involved, they also tend to know *who* the next speaker is. How turn-taking is regulated has been studied in the field of conversational analysis, where empirical studies of human conversations have been conducted. Sacks et al. (1974) propose that turn-taking is regulated by rules, which apply at a transition-relevance place (TRP). In short, the rule says that the current speaker may select the next speaker, who then is obliged to take the turn. if the current speaker neglects to do so, another speaker may self-select. If no one takes the turn, the current speaker may continue. What still has to be considered is individual differences among the participants. Some participants are more eager to take the turn than others, and also to hold the turn. The status of the participants, their role (for instance doctor-patient, friends, mother-child etc) in the situation and other contextual features in the situation may also affect how and if the rule apply. In some contexts, *silence* may also have to be interpreted, for instance when it occurs as the second turn of an adjacency pair (Jurafsky and Martin, 2000). An example for testing these variations has been presented by Jan and Traum (2005), who describe an algorithm for simulated turn-taking among background characters in a virtual world. In this case it is not important to analyse the actual information that is being exchanged, instead they focus on the "appearance of conversation and the patterns of interaction", which includes probabilities for **talkativenes** (wanting to talk), **transparency** (producing explicit positive and negative feedback and turn-claiming signals), **confidence**(interrupting and continuing to speak during simultaneous talk), **interactivity** (the mean length of turn segments between TRPs) and **verbosity** (continuing the turn after a TRP at which no one is self selected).

2.3 Characterization

In digital role-playing (adventure) games, such as the single player games Morrowind and Baldur's gate and the massively multiplayer Anarchy online, the player may set the personality of the player character in terms of for instance race (character class), gender, looks, skills and profession. These characteristics may have impact on how well the character may perform in specific situations in the game. In for instance Star Wars Galaxies, the player may also set the mood of the character, which is then displayed to other characters through text and facial animation (Eladhari, 2006). The main difference between the character representing a typical embodied SCA and character representing an in-game SCA lies in how the player relates to the character. In a game world, the rhetorical framing is constituted by *identification*, i.e. that the player to some degree (or completely) identifies him/herself with the character, or as Bartle (2003) put it: "Players play virtual worlds in order to be themselves". Another significant difference is that the characters of a game are placed in a fictive setting. The voice of a person is such a characteristic feature that adds to our full perception of that person. Apart from that, a personalized voice (perhaps chosen by the

player in the characterization phase), could be used in the game to identify the current speaker (in for instance multi-party dialogues). The character's personality and emotional state may be expressed using emotional speech synthesis and facial animations (see for instance Beskow et al. (2004); Bulut et al. (2002).

2.4 User/Player Variability

Different users will use different strategies in their interaction with the system and in speech-based system there are also variations in the speech to consider. In the latter case we talk about *speaker variability* which refers both to variabilities within a speaker, such as emotional and physical state, as well as across speakers, such as accent, dialect, vocal tract length, gender, and age. The former case refers to the individual user's strategies in approaching the system. In spoken dialogue systems, for instance, one strategy could be to use only isolated words in the interaction. In game systems, different player types can be identified and classified based on how they interact with the system and their motivation for playing the game (see for instance Bartle (1996, 2003)). We will not elaborate player types further in this paper, but for future studies, a classification of player types is required.

3 Summary

The aim of this paper has been to present a number of features that are required if we are to place a spoken conversational agent in a virtual game world. We chose to define typical SCAs as capable of handling task-oriented dialogues using the categorization proposed by Gustafson (2002) and compare those with *in-game* SCAs, capable of handling explorative and context-oriented dialogues in a *fictive* setting. A believable in-game conversation require incremental speech processing, rules for turn-taking, mixed-initiative and the possibility to have several participants (multi-party dialogue). For a typical SCA, the task(s) may accomplished without these features.

4 Conclusions

When introducing a new technology, or an "old" technology into a new field of application, the aim is to improve the application. When it comes to games, most development has been directed towards more realistic physics engines and 3D-graphics while little effort has been put into developing a more believable interaction with the characters in the game. Introducing more advanced speech technology is therefore not only a natural step in game development, it also adds new possibilities in game design, such as voice-based natural language conversation with non-player characters using for instance mobile phone or microphone. This paper has also shown that the requirements differ depending on whether we are developing a typical SCA or an in-game SCA. One major difference is

the importance of characterization, which is something needed in games and interactive stories, but not necessarily for other applications. The fact that games use fictive settings is also a significant difference, which also is one of the main differences between virtual game worlds and other virtual worlds.

References

James F. Allen, Donna K. Byron, Myroslava Dzikovska, George Ferguson, Lucian Galescu, and Amanda Stent. Towards conversational human—computer interaction. *AI Magazine*, 22(4):27–37, 2001. Available at: http://www.cs.rochester.edu/research/cisd/pubs/2001/allen-et-al-aimag2001.pdf.

R. Bartle. Hearts, clubs, diamonds and spades: Player who suit muds. *Journal of MUD Research 1*, 1, June 1996. Available at: http://www.mud.co.uk/richard/hcds.htm.

R. A. Bartle. *Designing Virtual Worlds*. New Riders, Indianapolis, Indiana, 2003.

J. Beskow, L. Cerrato, B. Granström, D. House, M. Nordenberg, M. Nordstrand, and G. Svanfeldt. Expressive animated agents for affective dialogue systems. In *Proc ADS'04*, 2004. Also available at: http://www.speech.kth.se/ctt/publications/papers04/ads04agents.pdf (2006-01-19).

S. Björk and J. Holopainen. *Patterns in Game Design*. Charles River Media, 2005.

Murtaza Bulut, Shrikanth Narayanan, and Ann Syrdal. Expressive speech synthesis using a concatenative synthesizer. In *Proceedings of ICSLP*, Denver, CO, 2002. Also available at: http://sail.usc.edu/publications/BulutNarayananSyrdal.pdf (2006-01-19.

M. Cavazza and M. Charles. Dialogue generation in character-based interactive storytelling. In *AAAI First Annual Artificial Intelligence and Interactive Digital Entertainment Conference*, Marina del Rey, California, USA, 2005. Available at: http://www-scm.tees.ac.uk/users/f.charles/publications/conferences/2005/AIIDE05CavazzaM.pdf (2005-12-14).

Jennifer Chu-Carroll. Evaluating automatic dialogue strategy adaptation for a spoken dialogue system. In *ANLP*, pages 202–209, 2000.

Nicolas Ducheneaut and Robert J. Moore. The social side of gaming: a study of interaction patterns in a massively multiplayer online game. In *Proceedings of the ACM Conference on Computer-Supported Cooperative Work (CSCW 2004)*, pages 360–369, New York, November 2004. ACM. Available at: http://www.parc.com/research/publications/files/5223.pdf (2005-12-14).

A. Egges, X. Zhang, S. Kshirsagar, and N Magnenat-Thalmann. Emotional communication with virtual humans, 2003. Available at: http://www.miralab.unige.ch/papers/161.pdf (2005—05—12).

M. Eladhari. The player's journey. In J. P. Williams and J. Heide Smith, editors, *Digital Gaming Cultures and Social Life*. McFarland Press, 2006. In Press.

J. Gratch, J. Rickel, E. André, J. Cassell, E. Petajan, and N. Badler. Creating interactive virtual humans: Some assembly required. Technical report, Workshop report, IEEE Intelligent Systems, 2002. Available at: http://www.cis.upenn.edu/ badler/papers/x4GEW.pdf (2005-05-11).

Joakim Gustafson. *Developing multimodal spoken dialogue systems. Empirical studies of spoken human-computer interaction*. PhD thesis, Department of Speech, Music and Hearing, KTH, Stockholm, 2002. Also available at: http://www.speech.kth.se/ctt/publications/.

Joakim Gustafson, Johan Boye, Morgan Fredriksson, Lasse Johanneson, and Jürgen Königsmann. Providing computer game characters with conversational abilities. In *Proceedings of Intelligent Virtual Agent (IVA05)*, Kos, Greece, 2005.

Haym Hirsh. Trends & controversies: Interactive fiction. *IEEE Intelligent Systems*, 13(6):12–21, 1998.

Dusan Jan and David R. Traum. Dialog simulation for background characters. In Themis Panayiotopoulos, Jonathan Gratch, Ruth Aylett, Daniel Ballin, Patrick Olivier, and Thomas Rist, editors, *IVA*, volume 3661 of *Lecture Notes in Computer Science*, pages 65–74. Springer, 2005. ISBN 3-540-28738-8.

Daniel Jurafsky and James H. Martin. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition*. Prentice-Hall, 2000.

C. A. Lindley. Trans-reality gaming. In *Proceedings of the Second Annual International Workshop Computer Game Design and Technology*, Liverpool John Moores University, UK, November 2004. Also available at: http://intranet.tii.se/components/results/files/WCGDT04reprint.pdf (2005-12-14).

M. Mateas and A. Stern. Natural language understanding in façade: Surface—text processing. In *Technologies for Interactive Digital Storytelling and Entertainment*, Darmstadt, Germany, June 2004.

M. Mateas and A. Stern. A behavior language for story-based believable agents. In K. Forbus and M. El-Nasr Seif, editors, *Working notes of Artificial Intelligence and Interactive Entertainment. AAAI Spring Symposium Series*. AAAI Press, Menlo Park, CA, 2002.

Michael F. McTear. Spoken dialogue technology: Enabling the conversational user interface. In *ACM Computing Surveys*, volume 34, pages 90–169, March 2002.

S. Mead, M. Cavazza, and F. Carles. Influential words: Natural language in interactive storytelling. In *10th International Conference on Human—Computer Interaction*, Crete, Greece, 2003.

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

G. Skantze. Exploring human error handling strategies: Implications for spoken dialogue systems. In *Proc ISCA Tutorial and Research Workshop on Error Handling in Spoken Dialogue Systems*, pages 71–76, 2003. Also available at: http://www.speech.kth.se/ctt/publications/ (2005-12-21).

W. Swartout, J. Gratch, R. Hill, E. Hovy, S. Marsella, J. Rickel, and D. Traum. Toward virtual humans. Working notes of the AAAI Fall symposium on Achieving Human—Level Intelligence through Integrated Systems and Research, 2004.

David Traum. *Advances in Agent Communication*, chapter Issues in multi-party dialogues, pages 201–211. Springer-Verlag LNAI 2922, 2004. Also available at: http://people.ict.usc.edu/ traum/Papers/multi.pdf.