ANGELICA
Choice of output modality in an embodied agent

Mariët Theune
Parlevink Language Engineering Group
Computer Science Department, University of Twente
PO Box 217, 7500 AE Enschede, The Netherlands
+31 53 489 4311
theune@cs.utwente.nl

1 Abstract

The ANGELICA project addresses the problem of modality choice in information presentation by embodied, human-like agents. The output modalities available to such agents include both language and various nonverbal signals such as pointing and gesturing. For each piece of information to be presented by the agent it must be decided whether it should be expressed using language, a nonverbal signal, or both. In the ANGELICA project a model of the different factors influencing this choice will be developed and integrated in a natural language generation system. The application domain is the presentation of route descriptions by an embodied agent in a 3D environment. Evaluation and testing form an integral part of the project. In particular, we will investigate the effect of different modality choices on the effectiveness and naturalness of the generated presentations and on the user's perception of the agent's personality.

1.1 Keywords

Language generation, embodied agents, modality choice

2 Introduction

In conversations between human speakers, speech is the main carrier of information, but nonverbal signals such as gestures and facial expressions also play an important role, providing additional information about the content and the structure of the discourse. With respect to their function, such nonverbal signals can be globally divided into three types. Signals of the first type reflect the structure of the ongoing discourse. For example, speakers may mark the introduction of a new discourse element by a quick hand movement, a nod, a raise of the eyebrows or a combination of these. The second type is that of deictic signals, which are used to indicate a specific, seen or unseen, discourse object. These signals usually take the form of a pointing gesture, but gaze direction, head nods and body movements are used as well. Signals of the third type are used to express part of the semantic content of a message, for instance by representing certain properties of objects and actions using gestures (hands forming a circular shape) or facial expressions (squeezed eyes symbolizing a small size [18]). In order to successfully engage in a lifelike interaction with a human user, an embodied conversational agent should be able to interpret the user's speech and nonverbal signals, and to respond with appropriate verbal and nonverbal behaviours of its own. The ANGELICA[1] project deals with the latter issue, addressing the generation of natural language combined with nonverbal signals.

Following Kendon and McNeill [12,17], we assume that language and nonverbal signals stem from the same conceptual source, and that neither can be regarded as being primary with respect to the other. This means that the generation of nonverbal signals cannot be separated from language generation in an embodied agent: both are part of the same information presentation task. For each piece of information to be presented by the agent, it must be decided whether to express it using language, a nonverbal signal, or both. In the ANGELICA project a model of the different factors that influence this choice will be developed, implemented, and tested. Before providing more details on ANGELICA, we briefly discuss some related work on natural language generation for embodied agents.

3 Language generation for embodied agents

Most embodied conversational agents or virtual presenters produce language using prefabricated scripts with canned utterances. There has been relatively little work on natural language generation (NLG) for embodied agents. Lester et al. [14] have worked on NLG in a pedagogical agent capable of generating simple deictic references, combining language with pointing gestures. André and Rist [1] have worked on NLG for virtual presenters, focusing on the problem of projecting different agent personalities.

So far, the issue of modality choice in information presentation by embodied agents has only been addressed extensively in the language generation work of Cassell et al. [4,5]. They concentrate on the division of labor between speech and gesture, making a distinction between ‘redundant’ gestures, which express the same semantic features as the accompanying speech, and ‘non-redundant’ gestures, which express information not expressed in speech. In [4], the choice between redundant and non-redundant gestures is related to information status: information that is somehow marked, e.g., because it is new

[1] A Natural-language Generator for Embodied, Lifelike Conversational Agents.

or contrastive, is expressed using both speech and gesture, whereas unmarked information is expressed using either speech or gesture. In [5], the choice between different possible combinations of gesture and speech is guided by information about the discourse context (which affects information status) and the communicative function of the utterance to be generated. The architecture for ‘embodied NLG’ described in [5] has been integrated in the REA agent, which presents descriptions of houses. No evaluation results have been published.

The focus of Cassell et al. on information status as guiding modality choice seems a valuable approach, and it will be interesting to see if it holds up in other domains and languages as well. However, other factors also seem to play a role in the distribution of information across different modalities, and these still remain to be investigated. In addition, as yet nothing is known about the effect that different distributions may have on the agent's audience. These issues will be addressed in the ANGELICA project.

4 The ANGELICA project

In this section we give a description of the ANGELICA project, starting with a global outline and then providing more details about our application domain and the central issues that will be addressed within the project.

4.1 Project outline

Within the ANGELICA project, to be carried out at the Parlevink Language Engineering Group at the University of Twente, we will develop a computational model of modality choice for information presentation by embodied agents. The modalities involved are spoken language and nonverbal signals. We focus on nonverbal signals with a deictic or a ‘content-bearing’ function, i.e., signals that are used to identify and describe objects and events playing a role in the discourse. In our domain, which is that of embodied route descriptions (section 4.2), these signals mostly have the form of broad arm and hand movements.

Our model of modality choice will be informed by video analysis of human utterances and by existing models such as those of Cassell et al. [4,5]. In addition to information status, we will also investigate the influence of other factors on modality choice (section 4.3). The model will be implemented as part of an information presentation component for embodied agents. Starting from a message specification, this component will produce natural language texts (in Dutch) that are automatically enriched with mark-up indicating the placement and global form of any accompanying nonverbal signals, as well as the placement of pitch accents and phrase boundaries. As a basis for the implementation we intend to use an existing language generation system for Dutch called LGM [20]. The LGM automatically determines the information status of the discourse items being expressed, which is relevant for the generation of both speech and nonverbal signals.

In order to use and test the modality choice model, the information presentation component will be integrated within a 3D embodied agent functioning as a virtual presenter. The other components of this agent's architecture, such as the modules guiding body movement and speech, will be based on previous and ongoing research in the Parlevink group (section 5).

4.2 Domain: embodied route descriptions

Ultimately, our aim is to develop a generally applicable framework for embodied information presentation, but within the ANGELICA project we will confine ourselves to the domain of route descriptions. This domain has been well-studied from an NLG perspective, but most of this work is purely language-based, generating either plain text [9, 11] or text with prosodic mark-up [21]. In the VITRA project, the generation of incremental route descriptions is supposedly combined with pointing, but no description is given of how this is done [16]. We are not aware of any work on the presentation of route descriptions by an embodied conversational agent.[2]

The route description domain has several properties which make it attractive for our project. Route description is a presentation task in which both deictic and content-bearing signals play a prominent role. Nonverbal signals in this domain typically have the form of broad arm and body movements, indicating a specific landmark or direction. Compared with more subtle (facial) signals, these broad movements are relatively easy to detect and to represent. Since route descriptions are in fact instructions on how to reach some location, the effectiveness of different modality choices may be measured in terms of the ease and speed with which users are able to reach their intended destination after having received a specific type of route description. Finally, the route description domain fits well within the research environment of the ANGELICA project. The Virtual Music Center (VMC, see Figure 1) developed at Parlevink is a 3D virtual building with halls, corridors and different floors. Since visitors to such 3D environments often experience navigation problems (see [23] for a discussion), the VMC is a natural environment for an embodied guide that presents route descriptions to visitors.

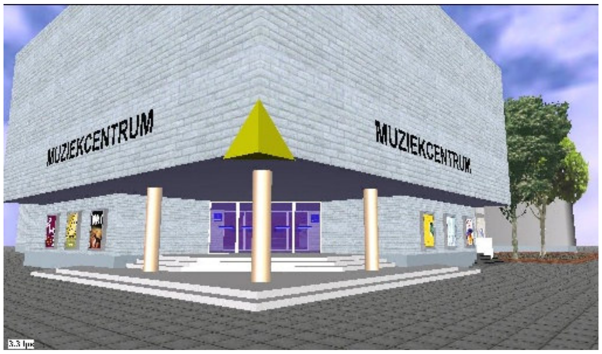

Figure 1. Outside view of the VMC.

[2] In the REAL project, an embodied agent shows users the way through a virtual environment, but this agent has no language generation capabilities except for limited object references [2].

So, route descriptions are a good starting point for our research. The initial version of our model will therefore be based on a video analysis of route descriptions by human speakers, and the implementation of a virtual guide presenting similar descriptions will serve as its test bed.

4.3 Research focus: modality choice

So far, research in natural language generation has been aimed primarily at uni-modal information presentation, where some underlying message is expressed using only natural language. However, when the message is to be expressed by an embodied agent an additional modality becomes available in the form of nonverbal signals. Not all parts of the underlying message can or should be expressed using nonverbal signals, so for each piece of information to be presented by the agent, at least the following questions must be answered:

- Is it *possible* to express it using a nonverbal signal?
- Is it *desirable* to express it using a nonverbal signal?
- Should it be expressed using *only* a nonverbal signal?

Below, these questions are discussed in more detail.

4.3.1 Constraints on the use of nonverbal signals

Not all types of information are equally suitable for expression using a nonverbal signal. Semantic features that can be easily visualized are the manner and direction of actions, and the shape, size and (relative) location of objects [12,17]. Abstract notions are less easy to convey using a nonverbal signal; here the use of some visual metaphor is required, such as depicting a physical container to represent a bearer of information (e.g., a film or story) [17]. When generating embodied information presentations, a simple way of checking whether a specific concept can be expressed using a nonverbal signal is to use a 'gesture dictionary' linking concepts to nonverbal signals [5,6]. Such a dictionary may be based on a (domain specific) inventory of nonverbal signals that have been actually produced by human speakers.

For deictic nonverbal signals the main constraint appears to be that the intended referent has a *location*, which may either be an actual, physical location or an abstract, relative position (e.g., a position on an invisible time line [17]). In route descriptions, most references are to physical objects with concrete locations, such as landmarks situated along the route. Therefore, in this domain most discourse objects can be indicated using a deictic nonverbal signal.

4.3.2 Selection preferences and production rate

Having identified those parts of the message that may be expressed using a nonverbal signal, some of these must be selected. Simply grasping all opportunities for generating a nonverbal signal is not an option, as it is likely to produce an unnatural effect, e.g., by giving the impression that the agent is "talking to a foreigner" [4].

Focusing for the moment on gestures, being the main kind of deictic or content-bearing nonverbal signal, we see that human speakers generally produce one gesture per clause [17], and that this gesture is usually located in that part of the clause where new information is presented [4,17]. Still, departures from this general rule are common [17], for instance when a speaker uses an additional gesture to mark 'old' information as contrastive [4]. In addition, there are several factors that can influence a speaker's overall gesture production rate, and that probably should be taken into account when attempting to generate lifelike embodied presentations. Among these are the following.

- As shown by Cohen [7] for the presentation of route descriptions, gesture rate (measured in gestures per second) increases with the *complexity* of the message being presented. (In [7], routes involving one or two choice points were considered simple; routes with more choice points were considered complex.)
- It is a common observation that gesture rate increases with the level of enthusiasm or *involvement* of the speaker [12]. Possibly related to this, Rimé and others found that the personality of speakers with a high gesture rate is more positively judged than that of speakers with a lower rate [19].
- Since gestures are at least partly produced for the recipient's benefit [7,13 and references cited therein], it is a reasonable assumption that gesture rate increases as the speaker attaches a higher *importance* to getting the message across; for example, in an educational setting the gesture rate may be higher than in social talk.

A possible way of dealing with the influence of such factors is to construct a hierarchy of 'gesture candidates', where the highest level is occupied by information that is most likely to be expressed using a gesture (e.g., new information that is relevant to the primary communicative goal, cf. [22]), and where the lowest level is occupied by the least eligible candidates for nonverbal expression (e.g., non-contrastive, discourse-old information). In a neutral situation, only information on the highest level is expressed in gesture, but influenced by factors like those mentioned above, candidates on lower levels may also be selected.

For non-gestural content-bearing and deictic signals like head or body movements it remains to be investigated how frequently they occur, where they are most often placed within an utterance, and how they are influenced by the abovementioned factors. Our general impression is that such signals are not very common, especially in the route description domain, and therefore they will not be in the main focus of our attention. Still, it seems reasonable to assume that as the complexity of the message or the involvement of the speaker increases, the production of non-gestural signals will increase as well.

4.3.3 Redundancy

Having selected those parts of the message that will be expressed using a nonverbal signal, the next decision to be made is whether this signal should be made redundant or non-redundant with speech. Bearing in mind that nonverbal signals run a higher risk of being missed (i.e., overlooked) by the recipient than speech, it seems that the choice between redundant or non-redundant signals is mainly determined by the deemed importance of the information to be expressed. In terms of the hierarchy mentioned in the previous section, we expect that items on the higher levels are more likely to be expressed redundantly than those on the lower levels of the hierarchy.

Another factor playing a role here is economy of expression. Some concepts are more easily expressed nonverbally than verbally. For instance, verbally describing an object is generally less efficient than simply pointing at it. Similarly, verbal descriptions of shapes or motions may be relatively complex in comparison to the corresponding gesture. In such cases, especially when the presentation is bound to certain time limits, the use of speech is less attractive, and the production of a (largely[3]) non-redundant visual signal will be preferred.

4.3.4 Remaining issues: form and timing

After having decided to generate a redundant or non-redundant nonverbal signal to express some part of the message to be presented, several issues remain to be addressed before an embodied presentation can be actually generated. Among these are the problems of properly synchronizing the nonverbal signals with speech, and of determining the actual form of the signal to be produced. Although these issues must definitely be addressed within the ANGELICA project, we will not go into them here.

4.4 Evaluation and testing

Evaluation and testing form an integral part of the ANGELICA project. The modality choice model will be evaluated through user experiments with the virtual guide, testing the naturalness and effectiveness of different types of generated descriptions, as well as their effect on the user's perception of the agent's personality. To facilitate testing and reuse of the guide's generation component, it will be set up so that the distribution of information across modalities can be varied using different parameters. We will conduct extensive user experiments, using different parameter settings and measuring and comparing the effect of the ensuing presentations. This way, we hope to find out if there is indeed an added value to generating nonverbal signals, when compared with a purely verbal route description. We expect that the generation of suitable nonverbal signals will increase the naturalness of the presentation and lead to a positive impression of the agent's personality (e.g., warm [19] and helpful). It remains to be seen whether the generation of nonverbal signals will make the presentations more effective, in the sense that visitors will be able to reach their intended destination quicker and more easily. There is experimental evidence indicating that nonverbal signals (i.e., gestures) help their recipients to understand and remember a message [13], but it has also been argued that nonverbal signals are generally not attended to, and that when they are, they are distracting and harmful to the processing of the message [8,19].

We also want to investigate the effect of additional factors such as those identified in section 4.3.2, for instance by increasing the production rate of nonverbal signals when generating descriptions of complex routes, and checking the consequences for perceived naturalness and effectiveness.

5 Related work at Parlevink

At the Parlevink language engineering group, which provides the research environment for the ANGELICA project, a 3D virtual theatre has been developed called the Virtual Music Center (VMC; see section 4.2). The VMC functions as a laboratory for research on virtual reality, agent-based software engineering, human-computer interaction, natural language processing and multi-modal dialogue. (See [10] for an overview.) The VMC is populated with several agents, both embodied and non-embodied, which provide various services to the visitor. One of them is a navigation agent that can compute the shortest route to the user's intended destination within the VMC [15]. This agent is not embodied. It has limited dialogue capabilities, but cannot give a verbal route description: it either presents the route on a map or moves the user through the VMC to his or her destination. Two embodied conversational agents developed in the VMC project are Karin, who can engage in a multi-modal dialogue about theatre performances and ticket reservations, and the virtual instructor Jacob (see Figure 2), who combines dialogue skills with the ability to move around and manipulate objects in his virtual environment.

Several research projects currently being carried out at Parlevink are aimed at improving the nonverbal skills of embodied agents. One project is aimed at the development of an agent body that is capable of more sophisticated movements than Jacob, while other projects focus on the generation of facial expressions [3], and on turn-taking behavior through gaze. The ANGELICA project is directly related to the projects sketched above, building on their results while adding natural language generation as a new, essential element. The virtual guide to be developed in the ANGELICA project will be an embodied version of the existing VMC navigation agent, extended with the new component for embodied language generation.

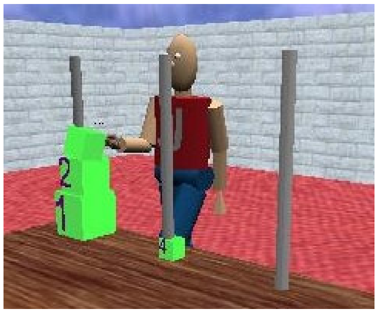

Figure 2. Jacob performing the 'Towers of Hanoi' task.

[3] Deictic references to objects usually co-occur with (at least) a pronoun or a demonstrative in speech. These verbal expressions function as 'syntactic placeholders' and depend on the deictic signal for their interpretation.

6 Final remarks

The ANGELICA project addresses an important issue in multi-modal information presentation: the choice of different output modalities for information presentation by an embodied agent. Language generation for embodied conversational agents is still a largely unexplored problem, and in the ANGELICA project, which is still in its preparatory phase, we hope to make a relevant contribution to this area.

REFERENCES

1. André, E. and T. Rist. Presenting through performing: on the use of multiple lifelike characters in knowledge-based presentation systems. *Second Int. Conference on Intelligent User Interfaces*, New Orleans, USA, 2000.
2. Baus, J., A. Butz and A. Krüger. Incorporating a virtual presenter in a resource adaptive navigational help system. *Workshop on Guiding Users through Interactive Experiences*, Paderborn, Germany, 2000.
3. Bui, T.D., D. Heylen, M. Poel and A. Nijholt. Generation of facial expressions from emotion using a fuzzy rule-based system. *14th Australian Joint Conference on Artificial Intelligence (AI'01),* Adelaide, Australia, 2001.
4. Cassell, J. and S. Prevost. Distribution of semantic features across speech and gesture. *Workshop on the Integration of Gesture in Language and Speech*, Wilmington, USA, 1996.
5. Cassell, J., M. Stone and H. Yan. Coordination and context-dependence in the generation of embodied conversation. *First Int. Conference on Natural Language Generation*, Mitzpe Ramon, Israel, 2000.
6. Cassell, J., H. Viljhálmsson and T. Bickmore. BEAT: The Behavior Expression Animation Toolkit. *28th Int. Conference on Computer Graphics and Interactive Techniques (SIGGRAPH),* Los Angeles, USA, 2001.
7. Cohen, A. The communicative functions of hand illustrators. *Journal of Communication 27* (1977), 54-63.
8. Cohen, A. The use of hand illustrators in direction-giving situations. In W. von Raffler-Engel (ed.), *Aspects of Nonverbal Communication*, Swets and Zeitlinger BV, Lisse, 1980.
9. Fraczak, L., G. Lapalme and M. Zock. Automatic generation of subway directions: Salience gradation as a factor for determining message and form. *Ninth Int. Workshop on Natural Language Generation*, Niagara-on-the-Lake, Canada, 1998.
10. Heylen, D. and A. Nijholt. Designing and implementing embodied agents: Learning from experience. *Agents 2001 Workshop on Multimodal Communication and Context in Embodied Agents*, Montreal, Canada, 2001.
11. Jokinen, K., H. Tanaka and A. Yokoo. Planning dialogue contributions with new information. *Ninth Int. Workshop on Natural Language Generation*, Niagara-on-the-Lake, Canada, 1998.
12. Kendon, A. Gesticulation and speech. In M.R. Key (ed), *The Relationship of Verbal and Nonverbal Communication*. Mouton, Den Haag, 1980.
13. Kendon, A. Do gestures communicate? A review. *Research on Language and Social Interaction 27*, 3 (1994), 175-200.
14. Lester, J., J. Voerman, S. Towns and C. Callaway. Deictic believability: Coordinating gesture, locomotion, and speech in lifelike pedagogical agents. *Applied Artificial Intelligence* 13, 4-5 (1999), 383-414.
15. Luin, J. van, A. Nijholt and R. op den Akker. Natural language navigation support in virtual reality. *Int. Conference on Augmented, Virtual Environments and Three-dimensional Imaging (ICAV3D).* Mykonos, Greece, 2001.
16. Maaß, W. How spatial information connects visual perception and natural language generation in dynamic environments: Towards a computational model. *Second Int. Conference on Spatial Information Theory (COSIT).* Vienna, Austria, 1995.
17. McNeill, D. *Hand and Mind: What Gestures reveal about Thought*. University of Chicago Press, Chicago, 1992.
18. Poggi, I., C. Pelachaud, and F. DeRosis. Eye communication in a conversational 3D synthetic agent. *AI communications 13,* 3 (2000), 169-182.
19. Rimé, B. and L. Schiaratura. Gesture and speech. In R. Feldman and B. Rimé (eds.), *Fundamentals of Nonverbal Behavior*, Cambridge University Press, Cambridge, 1991.
20. Theune, M., E. Klabbers, J.R. de Pijper, E. Krahmer and J. Odijk. From data to speech: A general approach. *Natural Language Engineering 7*, 1 (2001), 47-86.
21. Williams, S. and C. Watson. A profile of the discourse and intonational structures of route descriptions. *Sixth European Conference on Speech Communication and Technology (Eurospeech'99)*, Budapest, Hungary, 1999.
22. Yan, H. *Paired Speech and Gesture Generation in Embodied Conversational Agents*. Master's thesis, Media Lab, MIT, 2000.
23. Zwiers, J., B. van Dijk, A. Nijholt and R. op den Akker. Design issues for intelligent navigation and assistance in virtual environments. *Learning to Behave: Interacting Agents (TWLT 17),* Enschede, the Netherlands, 2000.