A CONVERSATIONAL AGENT TO NAVIGATE INTO MRI BRAIN IMAGES

Matthias LUDWIG(*), Alexandre LENOIR(+), & Pierre NUGUES(+)

(*) Fachhochschule Ravensburg-Weingarten
Doggenriedstrasse
D-88250 Weingarten, Germany

(+) Institut des Sciences de la Matière et du Rayonnement
6, boulevard du Maréchal Juin
F-14050 Caen, France

eMail: {mludwig, alex, pnugues}@greyc.ismra.fr

Tel: (33) 231-45-27-05 Fax: (33) 231-45-27-60

Correspondence

Pierre Nugues, ISMRA, 6, boulevard du Maréchal Juin, F-14050 Caen, France

pnugues@greyc.ismra.Fr

Abstract

This paper describes the prototype of a spoken conversational agent embedded within a simulation environment. This prototype accepts utterances from a user enabling him or her to navigate into a three-dimensional image of brain. The paper first describes what we can expect from such an interface in the communication quality between a user and represented in virtual worlds entities – artifacts –. Images have been obtained using magnetic resonance techniques. To enable a fast animation of images, they are reduced to surface elements. The paper describes how the images were processed and reconstructed. Then it describes the prototype's architecture which includes a speech recognition device together with a speech synthesizer. The system consists in a chart parser for spoken words; a semantic analyzer; a reference resolution system; a geometric reasoner, and a dialogue manager. The prototype has been implemented and has been demonstrated.

Keywords

Speech interface, Virtual reality, Simulation, Spoken navigation, Natural language processing, Reconstructed human body

Résumé

Ce papier décrit un prototype d'agent conversationnel incorporé dans un environnement de simulation. L'agent accepte les commandes orales d'un utilisateur lui permettant de naviguer à la voix dans l'image tridimensionnelle d'un cerveau. Le papier décrit dans un premier temps ce que nous pouvons attendre d'une telle interface dans la qualité de la communication entre l'utilisateur et les entités représentés dans les mondes virtuels – les artefacts –. Les images ont été obtenues par des techniques de résonance magnétique nucléaire. Elles ont été réduites à des éléments de surface pour permettre une animation rapide. Le papier décrit comment nous avons traité et reconstruit ces images. Ensuite, nous décrivons l'architecture du prototype qui comprend un système de reconnaissance et de synthèse vocale. Le système consiste dans un analyseur syntaxique fondé sur les charts, un analyseur sémantique, un résolveur de référence, un raisonneur géométrique et un gestionnaire de dialogue. Ce prototype à été implanté et nous en avons fait des démonstrations.

Mots clés

Interface vocale, Réalité virtuelle, Simulation, Navigation orale, Traitement automatique de la langue, Reconstruction du corps humain

Introduction

Medical education or surgery simulation require often three-dimensional images. These images are obtained from reconstruction techniques and can incorporate a representation of anatomical knowledge. Virtual reality techniques can be used to display 3-dimensional data and simulate manipulations. A spoken interaction – spoken dialogue – with the virtual reality system may facilitate interaction with entities of the virtual world because some 3-dimensional operations are easier to describe using voice than to carry out using a pointer. While speech interfaces are beginning to appear in virtual or simulation environments to ease interaction (Karlgren et al. 1995; Bolt 1980; Ball et al. 1995; Everett et al. 1995; Godéreaux et al. 1996), there are few medical implementations – none to the knowledge of the authors.

User interaction in simulation environments – virtual reality – is mostly done with more or less sophisticated pointing devices. These devices enable to move in horizontal and vertical planes and to rotate. They also enable to manipulate a specific object and to interact with it. However, navigating and interacting with entities in virtual worlds using devices such as mice, space balls, is a tricky issue for new users. Certain motions are difficult and a novice user can easily get seasick. One of the interaction difficulties is that the world is in 3 dimensions and that desired manipulations have to be described on a flat 2-dimensional screen.

Human-machine dialogue requires several relatively generic linguistic modules or devices such as speech recognition systems and speech synthesizers, syntactic parsers, semantic analyzers, and dialogue managers (Allen 1995). In a virtual environment, speech is only one mode of interaction – possibly a minor one – and some adaptations must be made to classical dialogue architectures. Notably, pointing devices must be integrated with speech.

Spoken interaction in a virtual environment requires to complement conventional pointing devices, to coordinate both means of interaction and to leave the user the choice – the initiative – of interacting means she/he wants to use. We found that in many situations the user prefer "to say it" rather than to "do it". However, it does not seem desirable to try to substitute completely these devices because it is sometimes easier to point at an object rather than to describe it in a verbal way.

Spoken interaction implies means to resolve deictic references that is coordinated with pointing devices and hence to reason about the geometry of the scene. Beside, the architecture must be complemented by an action manager that will realize the commands uttered by the user.

The Task

We investigated spoken interaction by designing scenarios to manipulate and navigate into magnetic resonance images of brain. Although the final goal our research is to serve as a medical application, the prototype has only been used for art performances. Our idea was to combine realistic images and dialogues to explore brain regions and their functions. The scenarios have been designed in cooperation with the art group *Das synthetische Mischgewebe*.

These scenarios have been limited to consider main regions of the brain such as the hemispheres, frontal lobes, parietal lobes, temporal lobes, the medulla, together with some specific sub-regions such as ventro-median regions, left temporal gyrus, amygdala, etc. These regions have been determined notably using Hanna Damasio's atlas (1995).

We restrained navigation to carry out linear motions and rotations relative to a designated object. The designation being uttered:

Go to the medulla and rotate around it

While it is not yet implement at the time we are writing this paper, the system should allow in the near future coordination of mouse designation and speech:

Go there and rotate around it (with a mouse pointing to an object)

The system can combine sequences of elementary motions. Manipulations of objects were designed as symmetric of navigation from the object coordinate system point of view. We set aside more complex motions such as tracking an object:

Follow the central sulcus

that would have required an extensive anatomical knowledge.

We also set as a requirement of the system to enable the user to query about his/her environment configuration. So that the system can provide certain descriptions of the virtual world. This includes position questions such as:

Where am I?

Or description of the location of other regions:

Where is <object>?

At the writing time of this paper such queries were not yet implemented.

Brain Images

One of the key points of virtual reality is real-time interaction. That is the system must respond and execute the commands immediately after they have been uttered. This implies to have a specific representation of brain images that enable fast computer processing. The representation is also required to have the different regions immediately perceptible. To meet these requirements, we represented the brain as a surface and we rendered the regions with colors and shading (Fig 1).

The 3D surface is build from an *in vivo* magnetic resonance imaging scanning provided by Cyceron research Center, Caen. This technique produces a 3-D image enclosed in a rectangular 3-D box. Image elements are cubical elements called voxels. The value of each voxel depends on the matter present at this point in the head of the subject scanned. A segmentation step first separates the brain from the head. Then the brain matter is segmented into three classes: cerebro-spinal fluid (CSF), gray matter (GM), and white matter (WM). The interface GM-WM is then tracked between gray and white matter with a surface tracking algorithm adapted from (Artzy et al. 1981).

The surface is extracted by a thresholding operation. The image is made of squares named surfels, which are the facets of the voxels (volume elements) of the volume image. Arbitrary points are chosen on the surface and associated to an arbitrary color. The colors are then diffused on the surface, and their intensity is modulated to take account that the area is inside a sulcus (darker), or at the top of a gyrus of the brain (lighter) in order to have a good perception of surface geometry (Lenoir et al. 1996). Finally we enhance contrast by a histogram equalization. We can use efficient displaying algorithm due to the surface model we use. The details of the surface are preserved and the colors produce a nice global looking effect (Fig. 1).

We implemented the image display on a Pentium powered computer without specialized hardware. The brain was sub-sampled two times in order to have a real time animation (the original surface had 170,000 surfels, and the surface of the sub-sampled surface had 40,000 surfels).

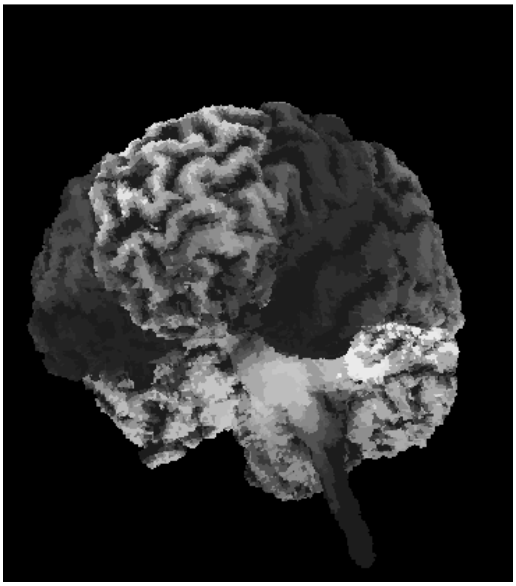

Figure 1 An image of the brain

System Architecture

The conversational system takes the form of a agent that is incorporated within entities of the virtual world – here only a brain. The system's overall structure is similar to that of many other interactive dialogue systems. It has been inspired by prototypes we implemented before (Nugues et al. 1993) and (Godéreaux 1994, Nugues et al. 1996). It features speech recognition and speech synthesis devices, a syntactic parser, semantic and dialogue modules. The system's architecture is also determined by the domain reasoner and the action manager.

The system capabilities were designed to be relatively specific. They concern navigation, manipulation, and some queries on the state of the world. The system assists the user within the world by responding positively to these commands. Understanding navigation commands also requires resolving references that occur in the conversation and to reason about the geometry of the world. Navigation or manipulations are completed by an action manager that carries out relatively continuous motion to bring the user where she/he wants to go or to animate the brain.

Speech recognition is carried out using the IBM's VoiceType commercial device. We have chosen this device because it can process French and can recognize other European languages with a vocabulary of up to 30,000 words. VoiceType is operating on isolated words – the speaker must pause between words – and is primarily intended for report dictation.

The whole prototype has been implemented on a single PC with the Windows 95 operating system. Each module is running independently as a thread of the application. The VoiceType *Speech Engine* is one of them.

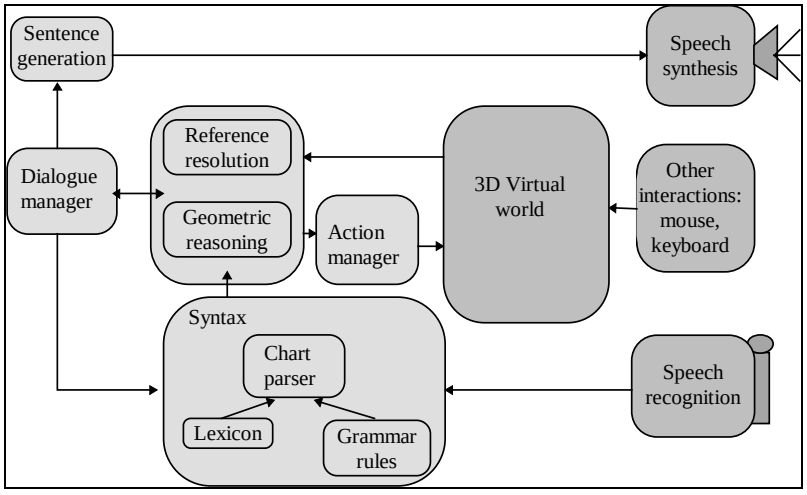

Figure 2 System Architecture

Syntactic Parsing

A chart parser is connected to the recognition device output and takes up the words. This chart (El Guedj et al. 1994) adopts a classical bottom-up algorithm with a dual syntactic formalism: It can operate using phrase-structure rules and a dependency formalism (Tesnière 1959). For this application, we chose the phrase-structure rule formalism. A constituent grammar was used to encode the lexicon – 800 distinct words – and relatively generic phrase-structure rules (Godéreaux et al. 1994).

Phrase-structure rules are rewriting the utterance structure using unification constraints and non terminal categories such as noun groups, verb groups, prepositional groups, determiner groups, adverb groups, adjective groups, etc. Rules were adapted to accept missing and unknown words. They include a large number of prepositional, adverbial, and demonstrative locutions that are ubiquitous in spoken language. The user segments her/his utterances using a "push-to-talk" scheme and signals the end of them by pressing a button.

Semantics Processing

Semantic interpretation considers navigation and manipulation commands and queries. It splits the utterance into clauses, and tags constituents from the chart parse tree with syntactic functions. Functions correspond to classical subject, object, or adjunct that are sub-classified using ontological categories. This stage also attaches modifying adverbs to their head words: verbs or other adverbs. Semantic annotation of verbs is related to the motion that is desired by the user and to space description. Considering previous experiments (Godéreaux et al. 1994) and lexical sources (Bescherelle 1980), we divided them into five main navigation categories:

1. go (*aller, avancer, entrer, monter, sortir*, etc.) corresponds to a change of location with a possible rotation of the user's point of view;
2. return (*revenir, retourner*, etc.) that is not implemented for the moment;
3. rotate (*regarder, se tourner, pivoter*, etc.) corresponds to the rotation of the user's point of view;
4. stop (*arrêter, stopper*, etc.)
5. continue (*continuer*)

As a result of this stage, each sentence is transformed in a list with as many items as there are clauses. Each clause is mapped to a structure whose members are the subject, verb group, and a list of complements. Verbs groups are annotated with a motion tag and packed with possible adverbs and clitic pronouns.

The logical form list is post-processed to relate it to a sequence of basic actions. These basic actions are compounds of translation and rotations. For instance *va devant le tronc* corresponds to:

1. the rotation of the user's point of view
2. the motion of the user to a specific view point associated with each object – here the medulla.

The reference resolution module de-indexes the sequence of action predicates resulting from the semantic interpretation. In this application, names are unambiguous which enables a relatively straightforward association.

Dialogue Processing

The dialogue module monitors the turn taking and the sequencing of the modules. It corresponds to getting the utterances, processing them, and executing them. The dialogue module manages the syntactic ambiguities by sequentially providing the semantic interpreter with the parse trees until it finds a correct one. It then passes the clause list to the reference resolution manager.

If an utterance corresponds to executable commands or to queries it can answer, the system will either acknowledge using a random positive message while doing them or answer the query. Otherwise, the dialogue manager rejects the utterance, indicating the cause. The natural language generator uses template messages and possibly selects a random one.

The Action Manager

The Action manager queries the Geometric reasoner to convert the referenced list of actions into a sequence of position coordinates. The reasoning is based on the verb category of each item of the action list. According to the category different kinds of actions are undertaken:

- go corresponds to a change of location and to a sequence of space positions. It can be preceded by a turn to have the user look to the object.
- turn corresponds to a rotation of the head.
- stop will stop the action
- continue will resume the action

The user is moved to his/her final location by transforming the geometric model of he brain using translations and rotations. This enable the user to move in a direction while looking in a different one. When going to a location a simple motion could implement a translation motion with a reasonable distance, but the user would loose the eye from the entity he/she probably wants to consider. In our prototype, the sight is directed on the main entity of the utterance. This makes the user feel more comfortable in the virtual world.

We associated each object with reference axes originating at its gravity center and view points to position the user relatively to it when the object is mentioned. This enable to take into account the different size of entities. A small entity will have view points closer than bigger ones.

The action are implemented by planning the action from small sub-motions. This generates a sub-motion list that are associated transformation matrices. Matrices are computed to reflect the view of the world – the brain form the user's position. The sub-motion interval can be adjusted to vary the speed.

Conclusion and Perspectives

We have presented a conversational agent that enables a user to carry out relatively complex motions within a virtual brain using voice. Our prototype consists in a commercial speech recognition device, together with a speech synthesis circuit. It relies on a modular architecture surface images of brain colored according anatomic regions. The entities of the prototype are to process syntax and semantics, together with dialogue and actions or

answers that are resulting from them. This prototype has been implemented on a Pentium class PC and demonstrated in art performances in France and abroad.

In the near future, we plan to adapt the prototype to other motions and to manipulations. We are also implementing more queries. In conclusion, we think that this kind of system can help users virtual environments and extend the diffusion of virtual reality in medical education.

A Dialogue Example

The action manager enables animations such as the sequence on Fig. 3. that correspond to the utterance:

Je voudrais voir le tronc. (I would like to see the medulla).

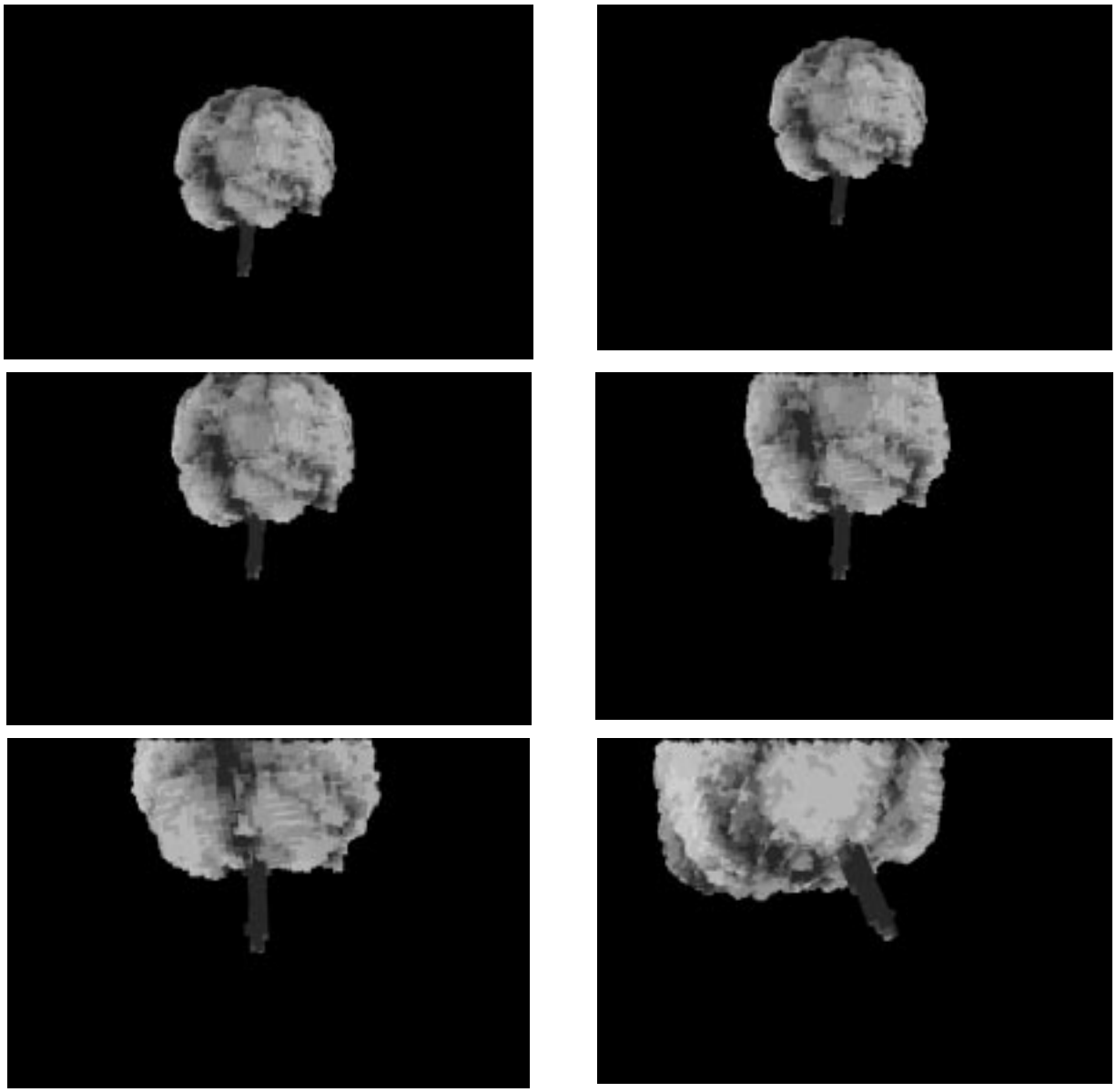

Figure 3 An animation Example

References

ALLEN, J.F.: *Natural Language Understanding*, Second edition, Benjamin/Cummings, 1995.

ARTZY E., FRIEDER G., HERMAN G.T.: The Theory, Design, Implementation and Evaluation of a Three-Dimensional Surface Detection Algorithm, *CGIP*, 15, pp. 1-24, 1981.

BALL, G. ET AL: Likelike Computer Characters: The Persona Project at Microsoft Research, in Software Agents, J. Bradshaw ed., MIT Press, To appear.

BESCHERELLE, *L'art de conjuguer*, Hatier, 1980.

BOLT, R.A.: Put That There: Voice and Gesture at the Graphic Interface, Computer Graphics, vol. 14, n° 3, pp. 262-270, 1980.

DAMASIO, H.: *Human Brain Anatomy in Computerized Images*, Oxford University Press, 1995.

EL GUEDJ, P.O. & NUGUES, P.: A chart parser to analyze large medical corpora, Proceedings of the 16th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, Baltimore, pp. 1404-1405, November 1994.

EVERETT S., WAUCHOPE K., PEREZ M.A.: A Natural Language Interface for Virtual Reality Systems, *Technical Report of the Navy Center for Artificial Intelligence*, US Naval Research Laboratory, Washington DC, 1995

GODÉREAUX, C., DIEBEL, K., EL-GUEDJ, P.O., REVOLTA, F. & NUGUES, P.: Interactive Spoken Dialogue Interface in Virtual Worlds, One-Day Conference on Linguistic Concepts and Methods in Computer-Supported Cooperative Work, London, November 1994, To appear Springer Verlag.

KARLGREN, J., BRETAN, I., FROST, N. & JONSSON, L.: Interaction Models, Reference, and Interactivity in Speech Interfaces to Virtual Environments, 2nd Eurographics Workshop, Monte Carlo, Darmstadt, Fraunhofer IGD, 1995.

LENOIR A., MALGOUYRES R., REVENU M.: Fast Computation of the Surface's Normals of a 3-D Discrete Object, to appear in *Lecture Notes in Computer Science*, Discrete Geometry for Computer Imagery '96, Proceedings, Springer, 1996.

NUGUES, P., GODÉREAUX, C., EL GUEDJ, P.O. & CAZENAVE, F.: Question answering in an Oral Dialogue System, In: Proceedings of the 15th Annual International Conference IEEE/Engineering in Medicine and Biology Society, Paris, vol. 2, pp. 590-591, 1993.

NUGUES, P., GODÉREAUX, C., EL GUEDJ, P.O. & REVOLTA F.: A Conversational Agent to Navigate in Virtual Worlds, in: Proceedings of the Eleventh Twente Workshop on Language Technology, LuperFoy S. Nijholt A., and Veldhuijzen van Zanten G. eds., pp. 23-33, 1996.

TESNIÈRE, L.: *Éléments de syntaxe structurale*, Klincksieck, 1959.