Automatic Speech Grammar generation during conceptual modelling of Virtual Environments

Lode Vanacken[1], Chris Raymaekers[1] and Karin Coninx[1]

(1) Hasselt University, Expertise Centre for Digital Media
and transnationale Universiteit Limburg
Wetenschapspark 2, B-3590 Diepenbeek, BELGIUM
E-mail: {lode.vanacken,chris.raymaekers,karin.coninx}@uhasselt.be

Abstract

Speech interfaces are becoming more and more popular as a means to interact with virtual environments but the development and integration of these interfaces is usually still ad-hoc, especially the speech grammar creation of the speech interface is a process commonly performed by hand. In this paper, we introduce an approach to automatically generate a speech grammar which is generated using semantic information. The semantic information is represented through ontologies and gathered from the conceptual modelling phase of the virtual environment application. The utterances of the user will be resolved using queries onto these ontologies such that the meaning of the utterance can be resolved. For validation purposes we augmented a city park designer with our approach. Informal tests validate our approach, because they reveal that users mainly use words represented in the semantic data, and therefore also words which are incorporated in the automatically generated speech grammar.

Keywords: User interfaces & interaction techniques, speech interfaces, conceptual modelling

1 Introduction

The realization of virtual environments (VE) is still a technical and time-consuming process. Indeed, developers are often restricted to the use of a low-level programming language to define the virtual world, the interaction of the user with virtual objects and the objects' behaviours within the virtual environment. Besides or in combination with advanced interactive tools, a possible solution is to explore conceptual modelling, which aims to ease the development process of virtual environment applications by defining different aspects of the application on a higher level (Cuppens & Coninx, 2005). With respect to the support of virtual environment realization through conceptual modelling, research has been focusing on scene and interaction development. Speech interfaces, more specifically command languages using speech grammars, have found their way as an interaction technique, but facilitating their development has not yet received much attention up to now.

Speech interfaces are increasingly being used in virtual environment applications since this way of interacting allows for more flexible and natural forms of interaction within a virtual environment. They have to recognize what a user utters and have to interpret these utterances in order to perform an action. As speech recognition is still limited, speech grammars are designed such that they can be used in a specific application with limited task domains, such as banking and travel services (Cohen, 1992). Given the fact that the system can recognize user utterances expressed in a restricted language, the user input needs to be interpreted to determine its effect

on the virtual world. Therefore the system has to perform reference resolution for which the required knowledge can be divided into ontological, linguistic and contextual knowledge (McGlashan, 1995). Ontological and linguistic knowledge will be generated by the designer when creating virtual environments using conceptual modelling. Contextual knowledge, on the other hand, is gathered at runtime by the user when interacting with the application. Therefore, contextual knowledge can be used as an addition to ontological and linguistic knowledge when building the speech grammar, but it is not a necessity and it is currently not considered in our approach.

In this paper, we will explain our approach to using conceptual modelling in order to automatically generate a speech grammar which can be combined with interaction techniques. Speech grammar generation thus becomes part of the conceptual design phase of the VE development. In the next section we will discuss related work, after which we will briefly describe the conceptual modelling approach we are using. In section 0 our approach for automatic speech grammar generation is discussed followed by a case study which has been augmented by our approach for validation purposes. Finally we conclude with a discussion, a conclusion and some future work.

2 Related Work

Speech interfaces are often used in combination with direct manipulation because speech alleviates some of the disadvantages of direct manipulation such as the difficulty to express quantities (Cohen, 1992). Quickset (Cohen, et al., 1997), FUSS (Gorniak & Roy, 2005) and (Tue Vo & Wood, 1996) are examples of such direct manipulation systems in which speech is combined with a pen based interface. For virtual environments some examples are (McGlashan, 1995; Muller, et al., 1998; Cernak & Sannier, 2002; Kaiser, et al., 2003), they are usually combined with some form of direct manipulation (e.g. gestures). All these speech interfaces use a speech grammar which is hand-made. Therefore, depending on the application a new speech grammar has to be created.

(Irawati, Calderón, & Ko, 2005; Irawati, Calderón, & Ko, 2006) use semantic virtual environment information which is divided into domain dependent and independent information, which is represented through ontologies. The authors claim that in their solution domain dependent information is specific to the application while their domain independent information remains static. However, in our opinion their domain independent information is domain dependent as it contains semantic information such as “wall” or “ball”. The generation of a speech grammar and the coupling with the rest of their framework is not discussed. (Martínez, 2004) proposed an augmentation of existing open file formats (VRML, x3d) with metadata such that all objects have unique identifiers. Using this technique Martinez was able to use them as semantic information combined with fuzzy logic for reference resolution with as main problem that the system has to contain correct fuzzy sets. Another reference resolution approach has been proposed by (Pfeiffer & Latoschik, 2004), they incorporate several parameters such as a common ground, features, naming and spatial references which are similar to semantic information represented in ontologies (ontological + linguistic knowledge). Besides the former, they also address contextual knowledge using the dynamic/competing frame of reference of (multiple) user(s).

(Otto, 2005) introduced a framework which combines the W3C Resource Description Framework (RDF) (W3C - RDF, 2007), with his world model which has a similar structure as VRML or x3d to incorporate semantic information. We will be using a similar notation which essentially is RDF, namely OWL (OWL, 2007).

Finally, (Conti, Ucelli, & De Amicis, 2006) created a semi-automatic tool which makes it possible for the application to understand the speech input of the user. During the coding

process of the application the coder adds extra semantic tags to the code which are during compilation processed such that they can be used by the framework for understanding the user.

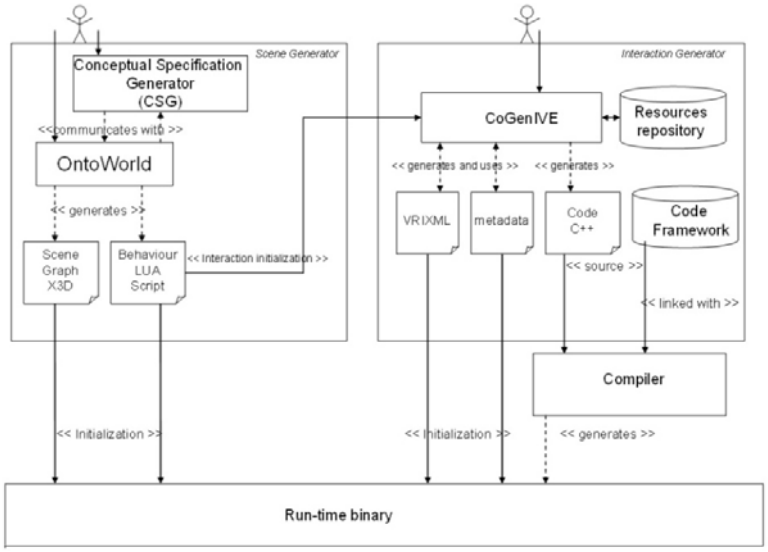

Figure 1 The VR-DeMo Approach (Coninx et al. 2006).

3 Conceptual Modelling

The VR-DeMo project (Coninx et al. 2006) provides us with inspiration for the conceptual modelling of the speech interface. The main aim of this project is to ease the development process of VE applications by high-level definitions of as much aspects of the application as possible. Currently, we concentrate on conceptual models and descriptions for the virtual world and the interaction, which are partly generated into source code and partly interpreted by the resulting applications (for an overview see Figure 1). In the context of this paper, we focus on the representation of the virtual world by means of ontologies. Possibly supported by an interactive tool, the designer creates an ontology containing domain knowledge to model the virtual world (ontological knowledge), and he defines a mapping from domain concepts to virtual objects and their interrelations (linguistic knowledge). The ontology is formulated in OWL. It is not only useful to express the structure of the virtual scene, but also to carry semantic data concerning the VE application. We will show in section 4.2 how we use this ontology to generate the speech grammar.

Currently, we make the ontology manually, but this activity could be supported by a tool such as OntoWorld that is being developed in the VR-DeMo project (Coninx et al. 2006).

An excerpt of such an ontology can be seen in Figure 2. It represents two concepts and two instances with their properties and relations. The domain concepts or contextual knowledge is represented as classes, properties and the possible relations between those concepts, while the mapping and linguistic knowledge are the instances of the classes, properties and their relations. For example a concept would be a *hotel* and its instance could be the *Hilton*. Note that the concepts, their instances and the relations between these are easy to generate and maintain in a tool.

4 Automatic Speech Grammar Generation

4.1 Process

Our approach to incorporate speech grammar generation in the conceptual design phase of the virtual environment consists of several steps: (1) The virtual world is modelled conceptually by

which semantic data is generated (see Section 3); (2) The semantic data is used to automatically generate a speech grammar; (3) This speech grammar is further annotated with synonyms using a lexical database of English: WordNet (WordNet, 2007). After generation, the speech grammar contains all pronounceable utterances specific to the virtual environment application. When spoken, these utterances still have to be resolved to an interaction or command. In order to interpret users' speech we need to perform reference resolution, in our case such an utterance will be translated to a query (SPARQL) which can be resolved using the semantic data generated earlier in OWL. The resulting information consists of the names of those virtual objects that satisfy the query, and these can consequently be passed on as input for an interaction in the virtual environment, such as selecting an object. An overview of the process can be found in Figure 3.

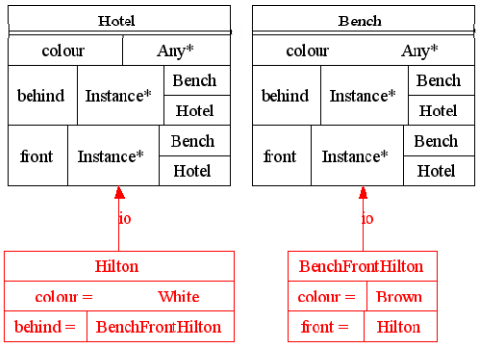

Figure 2 A Small Excerpt from an Ontology.

4.2 Generating the Speech Grammar

The process which generates the speech grammar receives as input the semantic data which has been generated during the conceptual modelling phase of the virtual world. The semantic data is represented as an ontology in OWL format, in Figure 2 an excerpt is illustrated. The generation of the speech grammar consists of several steps, each step uses different data of the ontology to finally become a pronounceable grammar.

The resulting speech grammar has the following structure:

- <command> <query>
- <o>What/How/Which is</o> [data-property] <o> of </o> <query>
- <query>
  - <o> all </o> [concept]/[instance]
  - [concept] [object relation] [concept]/[instance]
  - [data-property-value]* [concept]

Here <*command*> stands for a command for the application (e.g. "select"), <*o*> means optional and <*query*> stands for the part which indicates an object the user speaks about, this part needs to be resolved in a later stage and in order to show the possibilities we added one grammar rule which could be used in dialog systems: a question which can be asked to the application. All items between "[]" are types of semantic information modelled in the ontology. *[concept]* Stands for a concept or class in the ontology (e.g. "hotel"), *[instance]* is an instance of such a concept (e.g. "Hilton"), [object relation] interprets as a relation between concepts (e.g. "left of"), *[data-property]* is a data property of a concept (e.g. "colour") and finally *[data-property-*

value] is the instance of a data property (e.g. for colour "red") and which can also be repeated (indicated by "*"). In Figure 4 a part of the final resulting speech grammar for Microsoft Speech SDK 5.1 is illustrated, it has been created from the ontology in Figure 2.

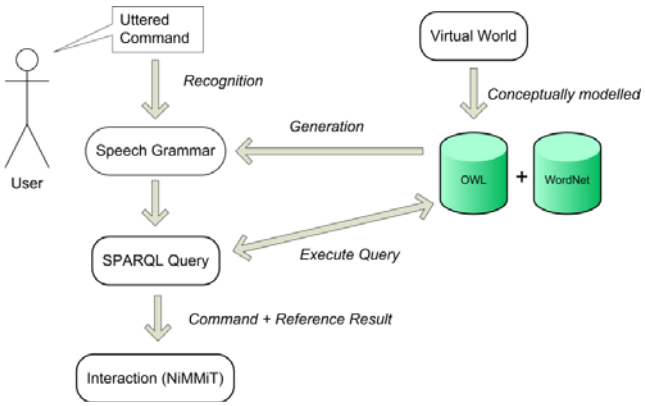

Figure 3 All Steps Performed during Speech Input.

```
<RULE NAME="CLASSES" TOPLEVEL="INACTIVE" >
    <L PROPNAME="CLASS">
        <P VALSTR="Bench">Bench</P>
        <P VALSTR="Hotel">hotel</P>
        <!-- Other Classes -->
    </L>
</RULE>

<RULE NAME="DATAPROPERTIESVALUES" TOPLEVEL="INACTIVE" >
    <L PROPNAME="DATAPROPERTYVALUE">
        <P VALSTR="colour,Brown">Brown</P>
        <P VALSTR="colour,White">White</P>
        <!-- Other Data Property Values -->
    </L>
</RULE>

    <!-- Other Semantic Data Rules -->
```

```
<RULE NAME="ALLQUERIES" TOPLEVEL="INACTIVE">
    <L>
        <P VALSTR="QUERYTYPE_CLASS_DATAPROPERTY">
            <P MAX="INF">
                <RULEREF NAME="DATAPROPERTIESVALUES"/>
            </P>
            <RULEREF NAME="CLASSES"/>
        </P>
        <!-- Other Queries -->
    </L>
</RULE>

<RULE NAME="MAINRULE" TOPLEVEL="ACTIVE">
    <L PROPNAME="QUERYTYPE">
        <P>
            <RULEREF NAME="COMMANDS" PROPNAME="INNERQUERYTYPE" />
            <RULEREF NAME="ALLQUERIES" PROPNAME="QUERYTYPE" />
        </P>
        <!-- Other Query Structures -->
    </L>
</RULE>
```

Figure 4 Parts of the Generated Speech Grammar.

For the structure of the speech grammar we based ourselves on other speech grammars used in related work and on a Wizard of Oz experiment performed by (Corradini & Cohen, 2002). We adopted an <action> <object> structure (Sharma, et al., 2000), which are respectively command and query. During each generation step we add WordNet synonyms to the speech grammar such that, if the user uses such a synonym, the system would still recognize the correct concept. Note that we could have added "–nyms" to the speech grammar such as hyper-, hypo-, holo- or meronyms, but we decided not to add any such extra words. First of all, it would enlarge the speech grammar and could make recognition worse and secondly if the designer would like the user to be able to use words which belong to any of those categories he should have incorporated this during the conceptual modelling phase of the virtual world. For example a bike has a steering wheel or peddles, these are meronyms but if not added by the designer they are probably of no use to the application, i.e. they cannot be manipulated or selected, and should therefore not be included.

4.3 Reference Resolution

If a speech command, uttered by the user, is recognized and thus is a valid construction in the grammar, then the final step has to be performed: reference resolution. Because our speech

grammar has a consistent structure we can use this structure to perform the reference resolution. As in our case the semantic information is represented in an OWL ontology we need some mechanism to query this ontology. We are using SPARQL (SPARQL, 2007) a query language for RDF, and so it can be used to pose queries at OWL ontologies. In order to easily perform these queries we use the w2p library (Vanderhulst, 2007).

The <*query*> part (see section 4.2) of the speech grammar is the only part transformed to a query; this transformation is relatively straightforward and consistent for all types of queries. In Figure 5 several transformations from the <*query*> part can be seen, other transformations are performed similarly.

The answer of the query is the result of the reference resolution and is used as input for the command which is the only remaining unprocessed part of the spoken utterance. Note that at the moment we do not incorporate contextual knowledge, meaning that we do not take into account the history of previous utterances or the position of the user in the virtual environment. Adding such knowledge could filter or interpret the results of the query even further. For example if similar virtual objects are found at different locations, the object at the same location, where the user is, will more likely be the object referenced to.

Figure 5 SPARQL queries: (a) query structure: [concept] [object relation] [concept]/ [instance] (b) example of (a): "bench left of fountain". (c) Query structure: [data-property-value] [concept] (d) example of (c): "Brown obelisk".*

5 Case Study: a "City Park Designer"

In order to assess the validity of our approach for generating a speech grammar, we augmented a conceptually modelled "City Park Designer" application with speech input as an interaction technique, and generated the speech grammar automatically. Using the interactive "City Park Designer", it is possible to design the park by positioning objects, such as buildings and statues. Furthermore, the designer can simulate behaviours, such as moving cars and shadows projected by the changing position of the sun. The park designer can use the automatically generated speech input commands to select/move/delete/add an object during the design of the city park. An example of such a command would be "select the bus left of the bus shelter" to refer to the object to be selected in the virtual scene. If the bus is of a specific brand, the user would also have been able to use this brand name. Besides using interspatial relations, objects and object-names, the user could also ask the application for the value of a data property of a certain object. The question "What is the colour of the bench left of the fountain?" is answered by the application with: "The colour is Brown". Figure 6 gives an impression of the virtual city park. Informal validation tests confirm our approach, because they reveal that users mainly use words represented in the semantic data, and therefore also words which are incorporated in the automatically generated speech grammar. Throughout this paper the examples used to illustrate the approach were generated from this case study.

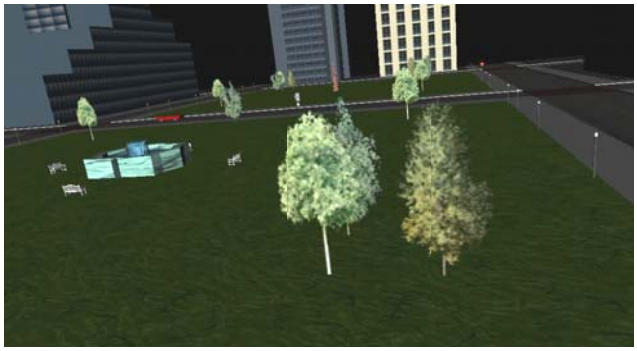

Figure 6 Impression of the Virtual City Park.

6 Conclusion and Future Work

In this paper we have introduced an approach to automatically generate a speech grammar using the semantic information generated during the conceptual modelling phase of the virtual environment application. The semantic information is represented using an OWL ontology and queries using SPARQL for reference resolution. We discussed the structure of the speech grammar and the conversion of this structure to the necessary SPARQL queries. Finally, we validated our approach by a case study in which we augmented a conceptually modelled "City Park Designer" with our approach. From this case study we could conclude that users used mainly words contained by the speech grammar during interaction with the application.
For future work we would like to incorporate contextual knowledge besides the ontological and linguistic knowledge which both come 'free' during the conceptual modelling phase. In order to further validate the approach a more formal user study could give more insights into the strengths and weaknesses of our approach.

Acknowledgements
Part of the research at EDM is funded by ERDF (European Regional Development Fund), the Flemish Government and the Flemish Interdisciplinary institute for BroadBand technology (IBBT). The VR-DeMo project (IWT 030284) is directly funded by the IWT, a Flemish subsidy organization. The authors would like to thank Geert Vanderhulst and the VUB-WISE partner for their help and advice.

7 References

Cernak, M., & Sannier, A. (2002). *Command Speech Interface to Virtual Reality Applications.* Virtual Reality Applications Center at Iowa State University of Science and Technology.

Cohen, P. R. (1992). The Role of Natural Language in a Multimodal Interface. *Proceedings of the Fifth ACM Symposium on User Interface Software and Technology*, (pp. 143-149). Monteray, CA, USA.

Cohen, P. R., Johnston, M., McGee, D., Oviatt, S., Pittman, J., Smith, I., Chen, L., Clow, J. (1997). QuickSet: multimodal interaction for simulation set-up and control. *Proceedings of the fifth conference on Applied natural language processing*, (pp. 20-24). Washington, DC.

Coninx, K., De Troyer, O., Raymaekers, C., & Kleinermann, F. (2006). VR-DeMo: a Tool-supported Approach Facilitating Flexible Development of Virtual Environments using Conceptual Modelling. *Virtual Concept 2006 (VC 06).* Cancun, Mexico.

Conti, G., Ucelli, G., & De Amicis, R. (2006). "Verba Volant Scripta Manent" a false axiom within virtual environments. A semi-automatic tool for retrieval of semantics understanding for speech-enabled VR applications. *Computers & Graphics (30):4* , pp. 619-628.

Corradini, A., & Cohen, P. R. (2002). On The Relationships Among Speech, Gestures, And Object Manipulation In Virtual Environments: Initial Evidence. *Proceedings of the International CLASS Workshop on Natural, Intelligent and Effective Interaction in Multimodal Dialogue Systems.* Copenhagen, Denmark.

Cuppens, E., & Coninx, K. (2005). CoGenIVE: Code Generation for Interactive Virtual Environments. *The Future of User Interface Design Tools, workshop of ACM Conference on Human Factors in Computing Systems (CHI 2005).* Portland, United States.

Gorniak, P., & Roy, D. (2005). Probabilistic Grounding of Situated Speech using Plan Recognition and Reference Resolution. *Proceedings of the 7th international conference on Multimodal interfaces*, (pp. 138-143). Torento, Italy .

Irawati, S., Calderón, D., & Ko, H. (2005). Semantic 3D Object Manipulation using Object Ontology in Multimodal Interaction Framework. *Proceedings of the 2005 international conference on Augmented tele-existence*, (pp. 35-39). Christchurch, New Zealand .

Irawati, S., Calderón, D., & Ko, H. (2006). Spatial Ontology for Semantic Integration in 3D Multimodal Interaction Framework. *ACM International Conference on Virtual Reality Continuum and Its Applications VRCIA*, (pp. 129-135).

Kaiser, E., Olwal, A., McGee, D., Benko, H., Corradini, A., Li, X., Cohen, P., Feiner, S. (2003). Mutual Dissambiguation of 3D Multimodal Interaction in Augmented and Virtual Reality. *In Proc. The Fifth International Conference on Multimodal Interfaces (ICMI 2003)*, (pp. 12–19). Vancouver, BC. Canada.

Martínez, J. I. (2004). *An Intelligent Guide for Virtual Environments with Fuzzy Queries and Flexible Management of Stories.* PhD Thesis at Universidad de Murcia.

McGlashan, S. (1995). Speech Interfaces to Virtual Reality. *Proceedings of 2nd International Workshop on Military Applications of Synthetic Environments and Virtual Reality.*

Muller, J., Krapichler, C., Nguyen, L. S., Englmeier, H. K., & Lang, M. (1998). Speech interaction in virtual reality. *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, (pp. 3757-3760).

Otto, K. A. (2005). The Semantics of Multi-user Virtual Environments. *Workshop towards Semantic Virtual Environments' (SVE 2005).* Villars, Switzerland.

OWL. (2007, July). Retrieved from OWL Web Ontology Language: http://www.w3.org/TR/owl-features/

Pfeiffer, T., & Latoschik, M. E. (2004). Resolving Object References in Multimodal Dialogues for Immersive Virtual Environments. *Proceedings of the IEEE VR2004*, (pp. 35-42). Chicago, USA.

Sharma, R., Zeller, M., Pavlovic, V. I., Huang, T. S., Lo, Z., Chu, S., Zhao, Y., Phillips, J. C., Schulten, K. (2000). Speech/gesture interface to a visual-computing environment. *IEEE Computer Graphics and Applications, 20* , pp. 29-37.

SPARQL. (2007, July). Retrieved from SPARQL Query Language for RDF.

Tue Vo, M., & Wood, C. (1996). Building an application framework for speech and pen input integration in multimodal learning interfaces. *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, (pp. 3545-3548). Atlanta, GA.

Vanderhulst, G. (2007, July). *w2p*. Retrieved from Web to Peer (W2P): http://research.edm.uhasselt.be/w2p/

W3C - RDF. (2007, July). Retrieved from http://www.w3.org/RDF/

WordNet. (2007, July). Retrieved from WordNet: http://wordnet.princeton.edu/