Computing in Civil Engineering:
Proceedings of the First Congress held in conjunction with A/E/C Systems 1994, Volume 2
Sponsored by the Committee on Coordination Outside ASCE of the Technical Council on Computer Practices of the American Society of Civil Engineers
Washington, DC, June 20-22, 1994

VOICE OPERATED INFORMATION SYSTEM (VOIS) FOR ELECTRONIC PAGES INFORMATION SYSTEM

Prasuna DVG Reddy[1], Ryuichi Kitamura[2] & Paul P. Jovanis[3]

ABSTRACT

This paper explains the on going work performed at UC Davis called Voice Operated Information System (VOIS) in the area of Advanced Travel Information Systems (ATIS). The aim of this work was to develop a habitable interface for the untrained user (driver), to investigate the degree to which dialogue control can be used to compensate for deficiencies in information systems interfaces. To give focus to our work, we have concentrated on electronic yellow pages context. However, the techniques developed are believed to be equally applicable to a wide range of other systems (pre-trip/en-route advice, route guidance system, etc.).

The dialogue controller is an independent unit with well-defined interfaces to the other system components. The Dialogue controller outputs a question to the speech output subsystem, and simultaneously outputs a set of syntax rules to the speech input system. These systems are very useful because there is no need for the driver to divert his concentration to the use the information system. Once this system is fully established, we are planning to use it as one of our prime user interface for all the prototype developments.

INTRODUCTION

The goal of this VOIS project is to provide transparent access to driver's information system. In order to achieve this goal, we needed to solve two major problems. First, in order to provide transparent access to database, for that we needed to build a framework which would allow us to monitor, model and translate to access information database without modifying it. Second, given specific requirements we needed to develop a methodology for translating information into voice (nonvisual) interfaces. This methodology is essentially the implementation of

[1]Research Associate, Inst. of Trans. Studies, Uni. of California, Davis.

[2]Professor, Kyoto University, Japan.

[3]Professor, Inst. of Trans. Studies, Uni. of California, Davis.

a hear-and-feel standard for this interface. Like a look-and-feel standard for graphical interfaces, a hear-and-feel standard provides a systematic presentation of nonvisual interfaces across applications.

In this paper, we describes the steps we have taken in the development of VOIS proto-type. In the following sections, we describe the system's over view. The second half of the paper describes the architecture we have constructed to provide this interface for yellow pages information database.

The other aspect of this system which is not explained in this paper is computer-generated voice which has been widely identified as a useful means of imparting information to drivers in an advanced driver information system (ADIS). A typical ADIS system may use voice to present navigation and traffic information without creating a visual distraction. In an Advanced Driver Information System, computer voice messages and non-verbal auditory signals provide a means of imparting information to the driver without creating a visual discretion. In the TravTek system [2], synthesized voice has been used extensively as a supplement to the visual display, providing route guidance instructions, navigation assistance, and traffic information. We will briefly characterize the current state of interactive interfaces with respect to the user's role. A thought-provoking, and much more detailed, examination is found in the [3] are especially relevant.

Terveen [5] observed that the model of human conversation may influence the structure of interfaces. Kant [1] also observed that the interfaces are dependent on type of application, type of users and the location characteristics. Sometimes high-end systems are not suitable for a type of users. So this work will gives a basic reactions for speech interface systems. A survey of commercially available speech recognition systems [4] has been conducted by display systems and simulation department of General motors gives the different speech recognition systems and their attributes.

NEED FOR VOICE INTERFACE

This investigation of speech dialog systems in vehicles is worthwhile for several reasons. Firstly, there is an increasing number of buttons and knobs that are necessary to control modern traffic information systems in the cars. But, as everyone knows, the optimal space for mounting these buttons and knobs is very limited. Secondly, the driver's capacity for multiple manual control is very limited. Too many manual controls may distract him from his primary task, that is handling and manoeuvreing the car. Thirdly, manual control of devices like route guidance systems is very inconvenient. Finally, we cannot ignore the fact that voice recognition systems are developed in laboratories all over the world. And, as history shows, once systems are developed, they will be introduced in our daily-life. So these systems are to be tested before they are install in the real life.

Before implementing actual voice control systems in cars it is important to answer the following questions:

1. what are the requirements for such systems from the user's point of view?

2. Are voice control systems which meet these requirements really useful for the driver?

These questions can be answered only when we do laboratory tests with proposed proto-type. This is the primary motivation for our VOIS project.

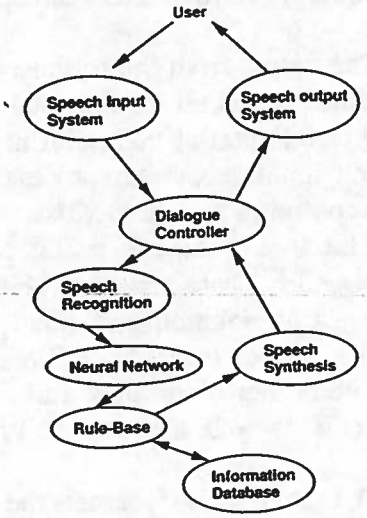

Figure 1: SYSTEM DESIGN

SYSTEM OVERVIEW

The VOIS system architecture is shown in Figure 1. As can be seen, the dialogue controller is an independent unit with well-defined interfaces to the other system components. A typical dialogue transaction cycle operates as follows. The Dialogue controller outputs a question to the speech output subsystem, and simultaneously outputs a set of questions rules to the speech input system (It is assumed that speech input system is located in the steering wheel. The probable locations for these units are not assumed yet. This depends on the type of vehicle and the information system. This paper only deals with prototype development and laboratory experiments which are planned to do on a micro-computer.). From the dialogue controller the input transfer to speech recognition. Speech recognition which is using trained neural network will try to recognize the input. If neural network system is not able to identify the word/words from the user then alternative system will be activated. For example, in the worst case dialogue may proceed as follows:

System: Where do you want to go to ?
User: I want to go to a mexican restaurant.

System: Sorry, where to ?
User: To a mexican restaurant.
System: Sorry, I still can't understand. Please state destination in one word only.
User: Restaurant.
System: Which type of restaurant you like to go ?
User: Mexican.
System: Sorry, I have problem of understanding you. Call xxx-xxxx for assistance.

The output from the neural network is sent to rule-base module where the rules define the subset of the total user input which the dialogue controller is prepared to interpret at that point in the dialogue. Using these rules as guidance, the speech input subsystem processes the user's response and returns it to the dialogue controller as a frame-like structure. At this point, the Dialogue controller accesses the local database, in this case, origin and destination locations, facility type and facility characteristics. Then it will communicates to central processing unit for getting solution and finally output the required information (voice & graphics) to driver (user) by VOIS synthesis and graphical display (optional). If dialogue controller didn't have full data to get information, then it will try to get from user. This cycle then repeats until users' query is fully established.

The above outline presents the broad framework in which we have addressed the overall structure of the system. In this paper we are not describing the internal architecture of voice synthesizer or recognizer. This paper is fully concentrated on developing a interface to information system through voice in the context of IVHS.

DESIGN METHODOLOGY

Most of the earlier attempts at specifying and implementing a dialog design have been based, whether explicitly or implicitly, on a finite state transition network possibly augmented by some adhoc mechanism for providing some measure of dialogue context. In our view such an approach is intractable in the long term since the number of possible paths that a dialogue may take will be very large. Thus, as the system evolves the dialogue controller software would rapidly become unmanageable. We have therefore adopted a data-driven methodology in which each item of information needed to construct a query, is represented by a frame (object).

Frames

A frame consists of data and methods (procedures) and it has two key properties (Figure 2). Firstly, the method of a frame may be invoked automatically when its data become needed (by the system) or when data is added by the system or user. Thus, it allows a dialogue system to be built in which the desired outcome of filling in all the required data slots is achieved without having to specify a precise order of dialogue events.

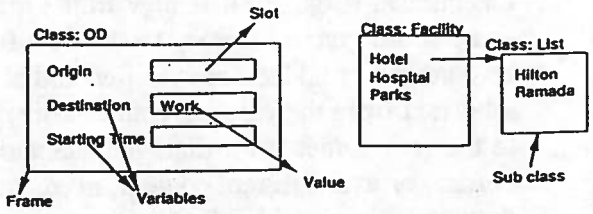

Figure 2: Textual Representation of Audio Frame

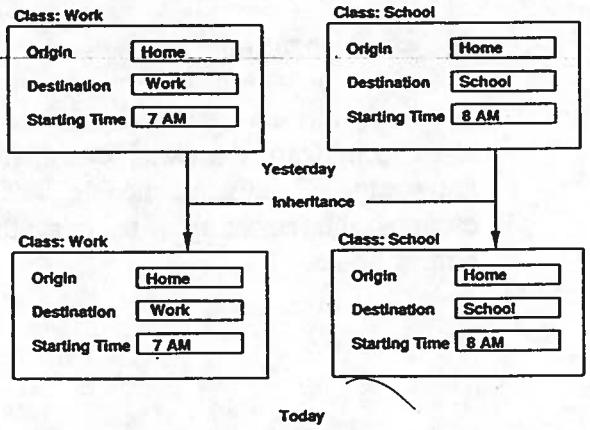

Figure 3: Textual Representation inherited mechanism of Audio Frame

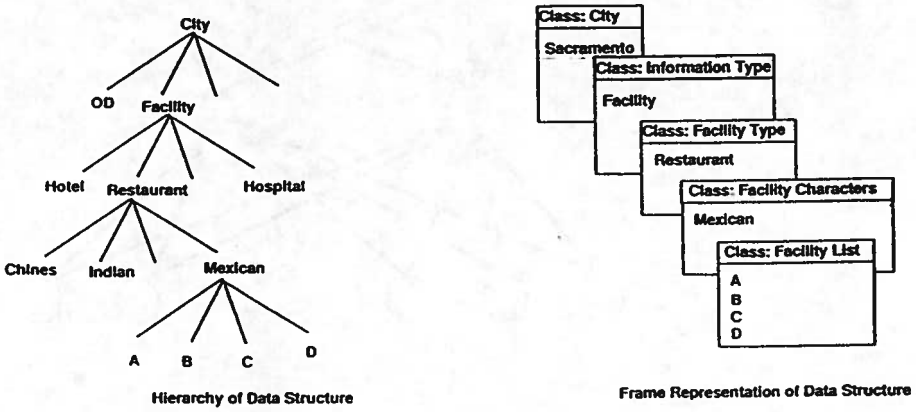

Figure 4: Data Structure & Frame Representation of VOIS

The second key property of a frame is that it has an associated inheritance mechanism (Figure 3). A new frame may be defined as being a specializations of some more general frame. In the new frame, only those properties which need to be changed or added are specified and all the remaining properties of the frame are inherited from the existing frame. This inheritance mechanism gives direct support to the requirements for dialogue adaptation. Adaption implies that something done before, or in a different context, needs modifying slightly. In the case of dialogues, the something could refer to a variety of things such as fixed origin and changing destination etc. For example, in the case of daily work trip, only starting or leaving time will change (Figure 3), all other items are same as previous day. Each of the information items required to construct a database query constitute valid topics of discussion and they are organized into a hierarchy reflecting the structure of the data (Figure 4).

Speech Recognition

Synthetic neural networks are an attraction to this problem, since they can learn to perform the classification from labeled training data and do not require knowledge of statistical models. The primary objective of this investigation is to establish the feasibility of using synthetic neural networks for the identification of human voice.

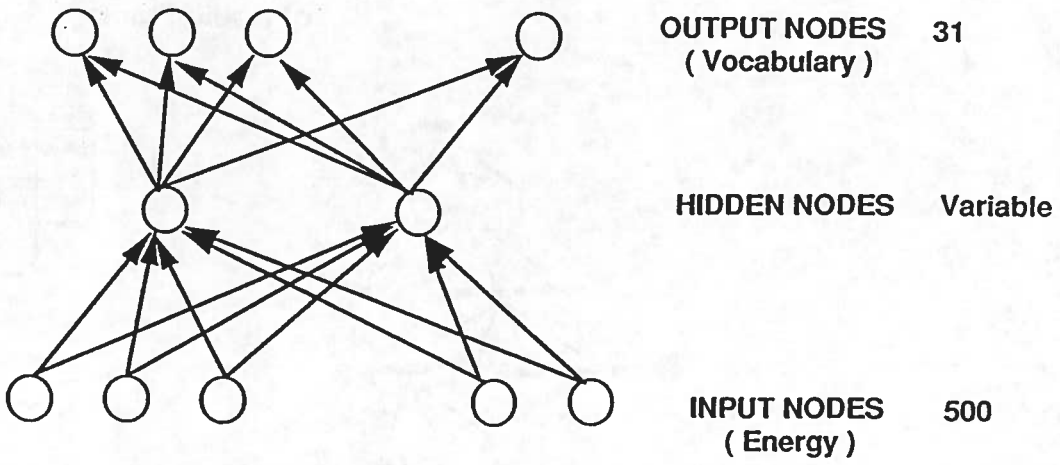

Figure 5: NEURAL NETWORK DESIGN

The neural network used in this study consists of the input layers (Figure 5), a hidden layer and the output layer. There are five hundred input node where each node represents speech energy at each time interval. The hidden layer have simulated nodes from 1/10th of total nodes to total number of nodes. Presently we have thirty one output nodes (total vocabulary). Each node represents a word. This network takes a large amount of time for training. The same network can be cut into small pieces to design into a step by step process. But driver has to listen to the output from the system and choose one out of them. For example, if system list hotels, parks, bus stops in the step by step process, driver has to say any one of them. In the case of other process, driver can say any word like 'Hotel Hilton', at any stage of system. Then the system will segment the sentence and take as two inputs. But in our recent studies, we found that step by step networks are showing very good results because the number of words to recognize for each network is very less compare to one single network. The other advantage in the step-by-step network is, if you extend the system (increase the number of words) you have to train only a small network. But in the case of single network you have to retrain the whole network from the beginning. During the training of the neural network, the output nodes are set to be '1', if that node is chosen or '0' otherwise. During the testing or prediction, the node is estimated to be chosen by the greatest value among the output nodes. With that the highest value from the output node should have minimum of 0.5. The learning model used is back-propagation method. Once the neural network recognized a word/words that will sent as a input to the rule-base.

PRELIMINARY RESULTS

These are the initial results observed when the system is tested by different people. These results are with respect of system design.

1. Vocal commands of drivers are often incomplete. These omissions do not follow normal grammatical or linguistic rules. They seem to reflect the user's naive assumptions about the actual state and the internal structure of the system under control. Omissions can occur at any point of the command structure. For example, instead of "Hotel Hilton" frequently the abbreviated command "Hilton or 'Hilton Hotel' is used. This shows users normally skips the or change the order.
2. Analog processes like search for a Hotel or opening next item are always broken down in discrete steps. From a theoretical point of view one would expect that drivers use commands like:"All Hotels, next, next,next, or Hotels and name". Instead they the command "All Hotels" and expect that the system list the Hotels.
3. Surrounding noise some times will disturb the recognition of words.
4. People have some initial problem in understanding the system. As time goes on, they will become more acquainted with the system.

FUTURE RESEARCH

To determine a catalog of requirements for voice recognition systems and to evaluate their impact on traffic safety a series of experiments should conduct using prototype. The first two experiments are planned to investigate the naive command vocabulary of drivers. Subjects are asked to perform various control actions in a car via verbal commands. One prominent expected result of these two experiments is that the use of nouns in verbal commands in cars is very consistent, even if the drivers have no experience with voice control devices. That means, a voice recognition system which is able to understand 3 or 4 synonyms for one device would react correctly in 95% of spontaneously given commands. However, the experiments may also show that the construction of user centered speech input systems are successful in cars and homes and may not successful at public places. Too many linguistic inconsistencies and too many changes in word order occurred. Therefore it was necessary to develop a model of ideal voice input commands and to modify this model on the basis of the obtained data.

REFERENCES

1. Kant., E. "Interactive Problem Solving Using Task Configuration and Control", IEEE Expert, 3(4) pp. 285-313, 1986.

2. Means., L. G., Carpenter, J. T., Szczublewshi, F. E., Fleischman, R. N., Dingus, T. A., and Krage, M. K., "Design of the Travtek Auditory Interface", General Motors Research, Michigan, 1992.

3. Norman., D. A.; and Draper, S. W., (ed.) "User Centered System Design: New Perspective on Human-Computer interaction", Hillsdale, NJ, 1986.

4. Schaeffer., M. S. " Speech Recognition Technology Survey", GM Research Report No. 89-27-77/G7690-001, 1989.

5. Terveen.; L. "Resources for Person-Computer Collaboration", Working notes from the AAAI Spring Symposium on knowledge-based Human-Computer Communication, pp.132-135.