Moving Toward Deployment:
Proceedings of the IVHS America 1994 Annual Meeting, Vol. 2
Atlanta, Georgia, April 17-20, 1994

VOICE OPERATED INFORMATION SYSTEM (VOIS) FOR DRIVER'S INFORMATION GUIDANCE SYSTEM

Prasuna DVG Reddy
Ryuichi Kitamura
Paul P. Jovanis
Institute of Transportation Studies
University of California

ABSTRACT

This paper describes work performed at UC Davis called Voice Operated Information System (VOIS) project in the area of Advanced Travel Information Systems (ATIS) as a part of IVHS. The principal aims of this work were to develop a habitable interface for the untrained user (driver), and 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 a pre-trip planning/en-route advice context. However, the techniques developed are believed to be equally applicable to a wide range of other information systems (electronic yellow pages, route guidance system, etc.).

In this work more emphasis is placed on media to interface with information systems. In other words the database is small and options are few in the information system. But the object of this study is to concentrate on the benefits and difficulties in using voice as a user interface media.

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 rules define the subset of the total user input language 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. These frames have information about user request. The Dialogue controller interprets the reply frame and the cycle then repeats until the user's query is fully established.

The above outline presents the broad framework in which we have addressed the dialogue controlled pre-trip planning information systems. These systems are very useful because there is no need for the driver to divert his concentration to use the information system. Once this system is fully established, we are planning to use it as one of our prime user interfaces for all the prototype developments.

INTRODUCTION

The goal of human-computer interface (HCI) is to provide a communication pathway between computer software and human users. The history of human-computer interface can be interpreted as the struggle to provide more meaningful and efficient communication between computers and humans. One important breakthrough in HCI was the development of user interfaces. In this context, interface provided is a graphical representation for data objects such as pre-trip, en-route and yellow pages information which are already available.

In this paper, all these informations are considered to be the part of pre-trip planning information. Unfortunately, present graphical user interfaces, or GUIs, have disenfranchised a percentage of the computing population. These user interfaces like click button driven and touch screen are all but complicated for commuters, because it is difficult for the driver to access information while driving. In other words the driver has to change his line of vision to the system to enter input or to read output.

This paper describes work performed at University of California at Davis called Voice Operated Information System (VOIS) project in the area of Advanced Travel Information Systems (ATIS). The principal aim of this work were to develop a habitable interface for the untrained user (driver), and to investigate the degree to which dialogue control can be used to compensate for deficiencies in recognition performance. It differs from other work in this area primarily in the extent to which their aims have been met within a single integrated framework.

To give focus to our work, we have concentrated on a pre-trip planning/en-route information database application and all of the examples in this paper will assume this scenario. However, the techniques developed

are believed to be equally applicable to a wide range of other information systems.

The goal of this work, called VOIS, is to provide transparent access to the 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 a database, we needed to build a framework which would allow us to monitor, model, translate and access an 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 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 describe the steps we have taken to solve these two problems. In the following section, we describe the system overview for the VOIS interface. We introduce the concept of audio GUIs and the abstract components of auditory interfaces in the context of pre-trip/en-route planning system. We also detail some of the techniques we are using to convey a range of interface attribute information via the auditory channel. The second half of the paper describes the architecture we have constructed to provide this interface transparently for an information database.

The other aspect of this system 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,[1] synthesized voice has been used extensively as a supplement to the visual display, providing route guidance instructions, navigation assistance, and traffic information. Special consideration has been given to strategies for maximizing the acceptability of synthesized voice to drivers.

CURRENT INTERFACES

User interfaces have experienced radical changes in recent years. Wide use of bit-mapped displays and mice have resulted in significant increases in the graphic presentation of information, and interfaces that require users to do practically no typing are well known. However, none of these techniques addresses the basic problem of how to raise the level of interaction to a dialogue. 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 found in Norman et. al. is especially relevant.[2]

Current systems tend to provide users with one of two distant roles in the interaction. In one role the system is in charge, and in the other the user is in control. When a user invokes a typical application, the program almost always initiates interactions, and the sequence in which the interactions occur is under the program's control. An advanced interface for an application may provide users with very easy modes of interaction (e.g., menus and active graphics), answer validation, and convenient ways to request canned, explanatory help messages about the meaning of the pending interaction. However, a user rarely has a means of questioning the interface about the implications of available choices or of gaining other information about future interactions that the program has planned -- much less exerting any influence over this scenario.

Terveen[3] observed that the model of human conversation may influence the structure of interfaces. Kant[4] 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 type of users. This work will give a basic reactions for speech interface systems. A survey of commercially available speech recognition systems[5] has been conducted by display systems and the 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 maneuvering 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. As history shows, once systems are developed, they will be introduced in our daily-life. These systems are to be tested before they are installed in real life.

Before implementing 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 this prototype. This is the primary motivation for our VOIS project.

SYSTEM OVERVIEW

Before explaining about VOIS system design, some of the important technical terms are explained to understand the design concepts.

Speaker Dependent Systems

Speaker dependent systems depend upon "enrollment" for their performance. Each speaker whose commands are to be recognized must pronounce each of the words in the vocabulary from once to several times. The speech patterns are statistically analyzed and recorded. When the speech recognition system is performing, spoken patterns are compared with those recorded by the appropriate speaker. These systems are very good in getting accurate matching rates.

Speaker Independent Systems

Speaker independent systems do not require enrollment of individual users. They operate with prerecorded patterns which are intended to be representative of the speech of the entire class of users or with patterns appropriate subclasses (e.g., men or women). These systems are a little complicated in achieving higher success matching rates. In spite of this problem, these systems are well suited in public places.

Total Vocabulary

Total vocabulary refers to the entire set of words that the speech recognition system will recognize. Usually this is limited by design or by available memory.

Active Vocabulary

Active vocabulary is the vocabulary that can be associated with a particular spoken signal by a speech recognition system. The smaller the active vocabulary, the fewer candidates there are to match, or mismatch, a spoken word. Active vocabulary can be defined by switch inputs to the system or by incorporation of a grammar which, in response to given spoken inputs, limits the recognition candidates for subsequent spoken inputs.

Isolated Word

Isolated word systems require a minimum pause, of the order of 100-200 ms, between spoken words. They use this period of no speech energy to establish the beginning and ending of words. However, a "word" need not be a single word of the language being spoken; it may be a phrase (utterance) limited by total spoken time to several seconds rather than by word length. Of course, the inter-word interval within the utterance will need to be less than the required minimum inter-utterance interval. Isolated word systems normalize the duration of the utterance to a standard length so that they are more or less immune to speaker rate.

Connected Word

Connected word systems determine word breaks by analyzing the speech fluctuations of the speech input, Their algorithms, called "Segmentation," are more sophisticated than those of isolated word systems, therefore they are more likely to be correct when they determine the beginning and end of words. Since misaligned word end points are responsible for the bulk of the word recognition errors made by the isolated word recognizer, connected word systems, operating in an isolated word mode, can be expected to improve the isolated word recognizer. In the connected word mode, performance will generally be degraded because the acoustic variation of words spoken in connected speech is greater than when they are spoken discretely, because of the influence of neighboring words on articulation.

Continuous Speech

Continuous speech systems, properly, should allow the speaker to converse with the speech recognition system in a "natural manner." This requires much more than just the ability to determine the ends of naturally spoken words. Actually, both isolated word systems and connected word systems can be made very sophisticated by the incorporation of a syntax. Use of a syntax allows many commands entity. Continuous speech systems of any complexity do not yet seem to be on the horizon. Their availability does not seem to depend upon improvements in speech recognition technology as such but rather on advances in artificial intelligence.

Performance Accuracy

Performance accuracy is measured by recognition rate. Properly, whatever is not recognized (unless it is a spurious sound or an improper utterance) should be counted as an error. But errors can be of different sorts with different consequences. Substitution errors occur

when one word is taken for another. They are usually of more consequence than non-recognition errors but, in an emergency, this need not be the case. Substitution errors and non-recognition errors have a reciprocal relationship; increasing standards for recognition tends to decrease the former and increase the latter. Another type of error occurs when a speech recognition system responds to noise or unauthorized words. The proper response to these events, if any, is non-recognition but substitution frequently occurs. When system tests are conducted, external noise are usually excluded and unauthorized words are not uttered, so error responses to these spurious inputs tend to be underestimated and system accuracies are usually overstated.

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 transfers 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:

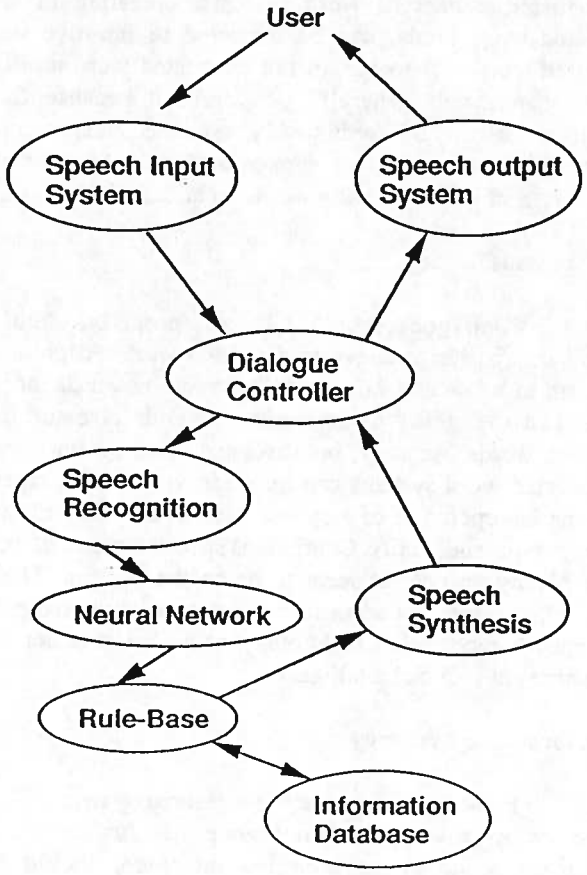


Figure 1. System Design

System: Where do you want to go to ?

User: I have to get to Sacramento this afternoon.

System: Sorry, where to ?

User: To Sacramento.

System: Sorry, I still can't understand you. Please state destination. One word only please.

User: Sacramento.

System: Sorry, I still can't understand you. Please call xxx-xxxx for operator 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. These reply frames are effectively just phrase structure trees with semantically redundant branches removed. The dialogue controller interprets the reply frame and the cycle then repeats until the user's query is fully established. At this point, the dialogue controller accesses the local database, in this case, origin and destination locations which drivers commonly use, travel timings and other route options. Then it will communicate with the central processing unit for getting a solution and finally output the required information (voice & graphics) to driver (user) by VOIS synthesis and graphical display (optional).

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 an interface to the 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 ad hoc 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). First, 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.

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 specialization 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

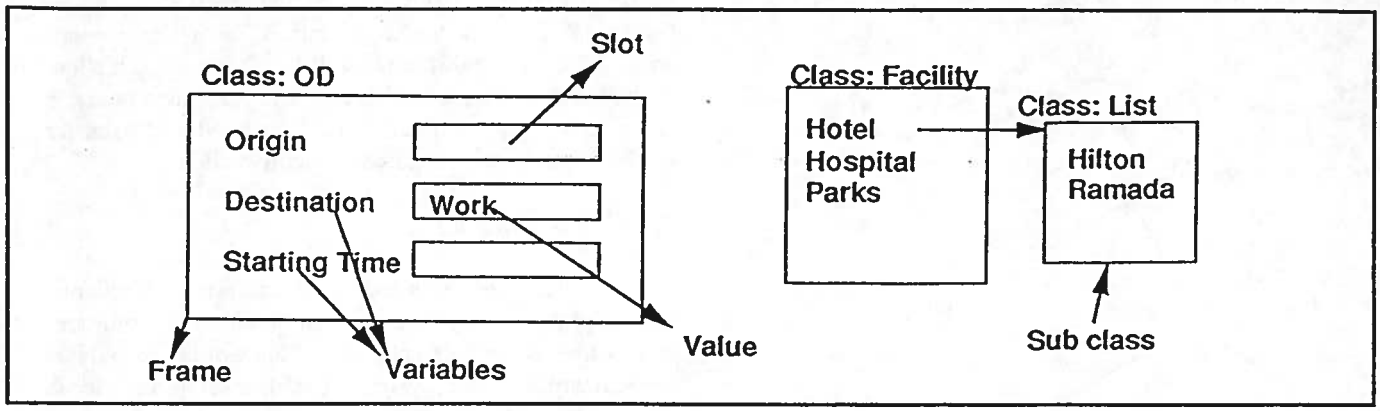


Figure 2. Textural Representation of Audio Frame

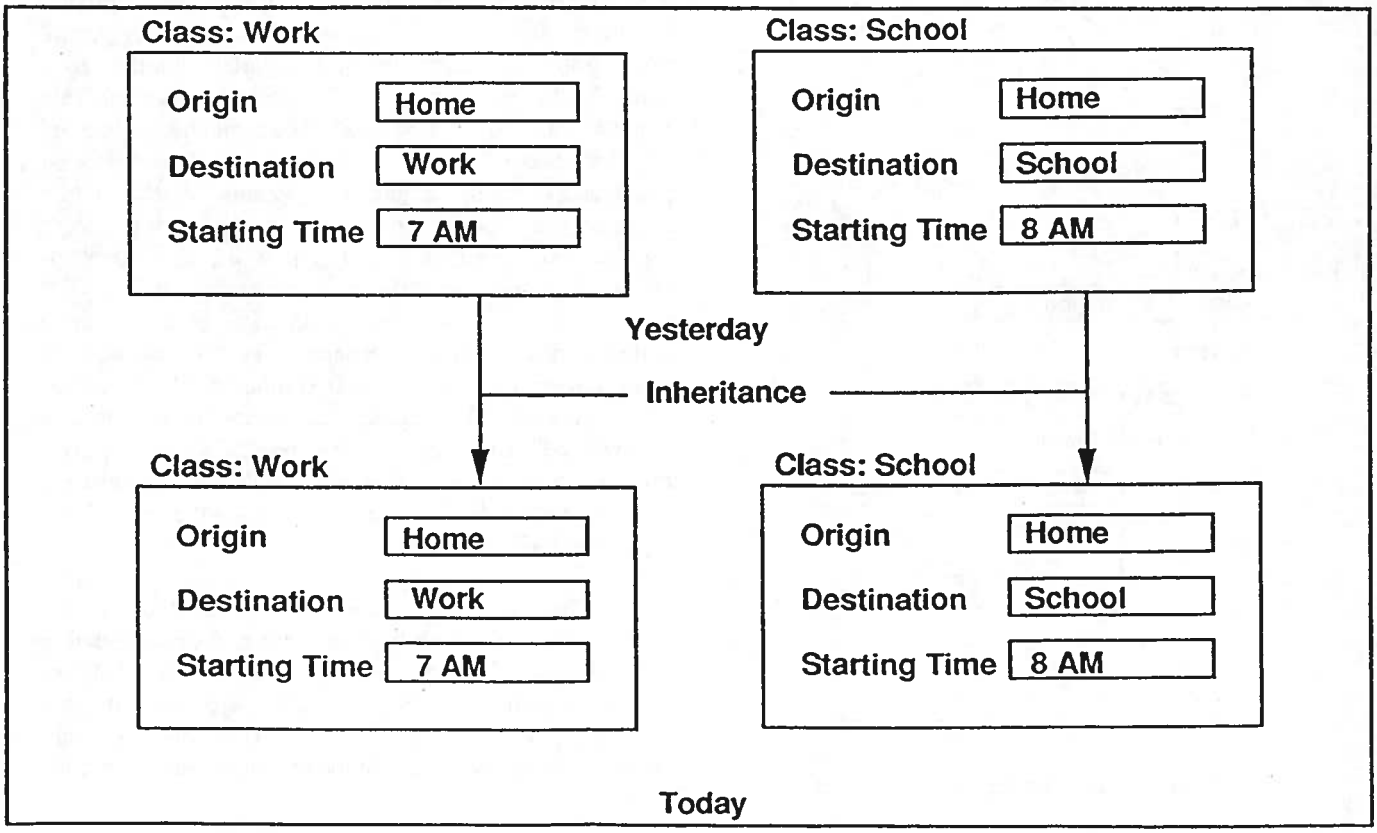
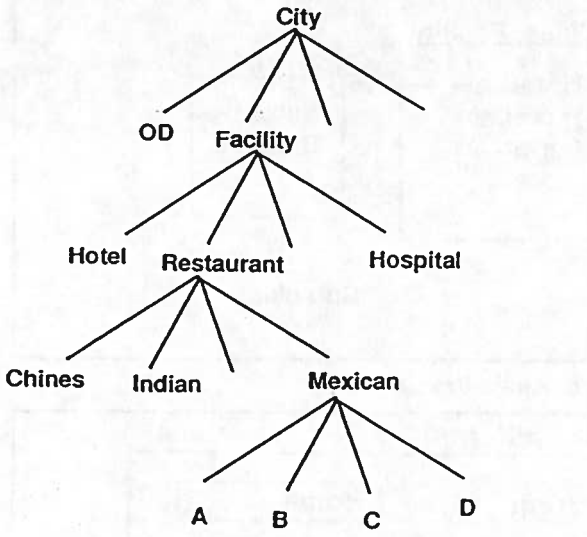


Figure 3. Textural Representation Inherited Mechanism of Audio Frame

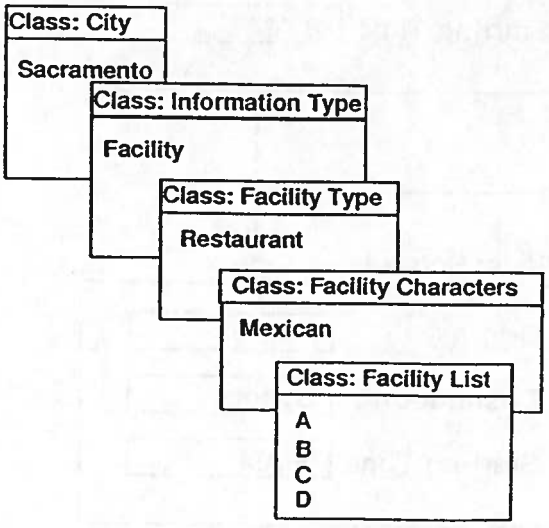
different context, needs modifying slightly. In the case of dialogues, the something could refer to a variety of things such as fixed origin and changing destination etc. For example, in the case of daily work trip, only starting or leaving time will change (Figure 3), all other items are the same as the 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. For example, Figure 4 shows the default structure of the initial query in the VOIS pre-trip planning application.

Figure 4. Data Structure & Frame Representation of VOIS



Hierarchy of Data Structure



Frame Representation of Data Structure

The hierarchical structure provides the primary context in which each question/answer cycle operates. Each node in this structure is represented by a frame. The methods of these frames are responsible only for maintaining the integrity of the data structure. For example, the origin frame is programmed with a suitable default leaving place (daily work trip) and the where frame's "if-added" method will check that the destination is the same as the origin. It means, if it is a daily regular trip, all the data is inherited by the previous day frame. If the starting time changed from yesterday, then only that field value will change. This way the required time to collect the data can be reduced significantly.

Since each node of the data hierarchy is also a potential dialogue topic, a second kind of frame called "dialogue" frame is attached to each node. In the simplest case, these dialogue frames will be a general question/ answer dialogue frame which asks an appropriate question when its associated node becomes needed and asks for confirmation when data is subsequently added.

Speech Recognition

Machine speech recognizer performs a frequency and amplitude analysis of spoken signals and compares them with stored representations of the words and phrases recognizable by the system. The first step in speech recognition is feature extraction. It represents the time-amplitude-frequency characteristics of the signal digitally, reducing data flow and analysis to manageable proportions. The next step is comparison of the input signal with the various stored reference patterns. This requires time synchronization of input signals and stored patterns, one of the most difficult speech recognition problems. Usually alignment methods establish the beginning and end of vocabulary entries using energy criteria: the beginning is signaled by the onset of significant speech energy and the termination by its dropping below some threshold for some time. The input pattern is now ready for comparison. For each vocabulary entry, a measure of similarity is computed. The reference word yielding the highest similarity rating is the "recognized" word except that the recognizer may apply a rejection criterion (more than a threshold value) to prevent incorrect recognitions. If this criterion is not achieved, the input word will be rejected.

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 a human voice.

Here, we followed two approaches in using this system. The first method applied to single user (speaker dependent system), which may be installed in his car or at home. A second method is applied to multi users (speaker independent systems), which is installed in public places like airports, roadside kiosks, etc.

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 nodes where each node represents speech energy at each time interval. The hidden layers have simulated nodes from 1/10th of total nodes to total number of nodes.

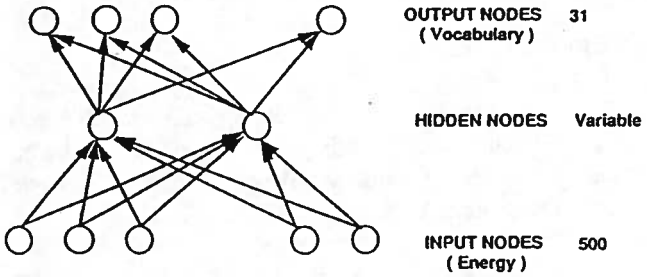


Figure 5. Neural Network Design

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 the driver has to hear the output from the system and choose one out of them. For example, if the system lists hotels, parks, and bus stops in the step-by-step process, the driver has to say any one of them. In the case of the other process, the driver can say any word like 'Hotel Hilton', at any stage of the system. Then the system will segment the sentence and take it 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 less compared 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 a single network you have to retrain the whole network. 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 a minimum of 0.5. The learning model used is back-propagation method. Once the neural network recognizes a word/words, that will be sent as an input to the rule-base.

Rule-Base

Once the speech recognition shorts out the word spoken by the driver, then the next step is to process the word. The element of work is taken care of by a rule-base module. This module contains a large set of rules called knowledge-base, which checks what action has to be taken depending on the input from the neural network. The final system will have a multiple input from different neural network (step-by-step) and store in a frame. This 'frame' is an input to a rule base which has full information about the request made by the driver (origin, destination, time, etc.).

The essence of the frame is that it is a module of knowledge about a situation, a phenomenon, or an object. Frames can be used for representing rather complicated concepts. They can contain various possible actions as well as conditions for the applying these actions. Besides these frames contain slots -- places to put concrete values or knowledge about other objects which can vary and are related to the concept. These slots are like the formal parameters of micro definition or like the parameterized data types. These frames check with an information systems database and come up with results, then transfer that results frame to a voice synthesizer and/or display unit. If there is insufficient data, then rule-base will send the same frame to speech synthesizer requesting for additional information.

Voice Synthesizer

This is the simple part. This module reads the frame/text input from the rule-base and using synthesized voice, the system will convey the message to the driver. This module has an extra feature other than just delivering speech, called repeat voice function. This function enables the driver to reply to the most recently spoken voice message.

Display Unit

This unit has a small graphic utility to display some graphics. This helps the driver to locate where he is and where he wants to go. Sometimes, if the driver does not understand the synthesized voice, this display terminal will help in conveying information to the driver.

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 that users normally skip words or change the order.

2. Analog processes, like searching for a Hotel or opening the 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 get the command "All Hotels" and expect that the system will list all the hotels for them.

3. Surrounding noise sometimes 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 be should conducted 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 three or four 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 be successful in 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. 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.

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

3. Terveen, L. "Resources for Person-Computer Collaboration," Working notes from the AAAI Spring Symposium on Knowledge-Based Human-Computer Communication, pp.132-135.

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

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