The Back Seat Driver: Real Time Spoken Driving Instructions

James Raymond Davis and Christopher M. Schmandt
The Media Laboratory
Massachusetts Institute of Technology
Voice: (617)-253-0314
Fax: (617)-258-6264

Abstract

The Back Seat Driver is an automobile navigation aid which uses synthetic speech to give driving instructions in real time to the driver of a car. The advantage of speech over visual aids is that it leaves the driver's eyes free for driving, however it also poses special problems. This paper describes the strategies employed by the Back Seat Driver to successfully use speech. We hope this paper will persuade you of the value of speech in driving directions.

Introduction

The Back Seat Driver uses synthetic speech to give driving instructions in real time to the driver of a car. Speech is the only output channel it uses. There are no graphics. This paper discusses the advantages and problems arising from our exclusive use of speech to provide directions. The first section presents a brief overview of the Back Seat Driver. The second section describes the linguistic abilities of the Back Seat Driver. The final section describes the problems we have encountered because of our exclusive use of speech, and how we have overcome them.

System Overview

The architecture of the Back Seat Driver is shown in figure 1. At the center of the Back Seat Driver is the map database. The street map represents two ways in which streets can be connected: *physical* connectivity means it is physically possible to drive from one segment to another, and *legal* connectivity means it is lawful to do so. Legal connectivity is obviously needed to find legal routes, and physical connectivity for correctly describing intersections. The street map also includes traffic lights, stop signs, the number of lanes, and the location of all gas stations. These features are useful for both route finding (since, e.g. fast routes should avoid traffic lights) and for descriptions. The location system (supplied by the project sponsor, NEC) determines the current position of the vehicle by dead reckoning and map matching. It is further described in [3]. The driver gives the Back Seat Driver a destination by entering an address on a keyboard. Using this map, the route finder can find the shortest route, the simplest one, or the one most easily followed, depending on the driver's preference.

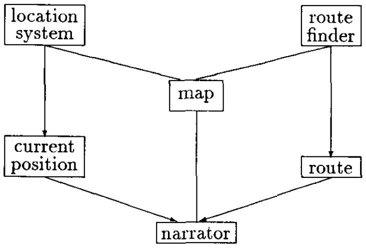

Figure 1: Back Seat Driver components

The *narrator* is the subject of this paper. It generates instructions spoken by a speech synthesizer (a Dectalk). The narrator follows the driver's progress along the route. It decides what to say by comparing the current position against the map. The system follows the driver's progress, giving each instruction just when needed. If the time between instructions is long, the program gives the instruction twice, first in a detail, and later in a brief form. When not otherwise occupied, the system may deliver voice mail messages, weather reports, or commentary about the route. If the driver makes a mistake the system automatically finds an alternate route and continues.

The system has been running in prototype form since April 1989. It has been successfully used by drivers who have never driven in Boston. A somewhat longer description of the system appears in [4]. A complete description of the system appears in [1].

Linguistic Abilities

In designing the Back Seat Driver we chose to use speech as the sole means of providing driving instructions for two reasons. First, we believe that the driver's eyes are already employed watching traffic, and best left undisturbed. Second, we know that the alternative (map displays) will not work for those people who have difficulty reading maps[5]. We were also influenced by an experiment on route following which compared spoken instructions with paper maps[6]. Subjects who heard spoken directions did better than those with maps, and also better than those with *both* sources of guidance. Although this experiment does not compare real time speech to real time maps, it does suggest that spoken directions might be easier to follow than visual directions.

Classifying Actions

Based on a study of how people naturally give spoken driving instructions, we developed a taxonomy of intersection types (Figure 2). This taxonomy is necessary in order to describe an intersection in the same way that a person would. For example, people talk about a "T" turn differently than a "fork" (or "Y") in the road. It is important that instructions match people's perceptions of the the world they see.

The proper classification of an intersection depends upon the topology (how many streets are at an intersection), the geometry (the angles among them), and the types of roads involved. For instance, the difference between the "T" and "fork" mentioned above is one of geometry, not topology (figure 3), and the difference between a "fork" and an exit from a highway is that one of the two roads in the "Y" of the exit is much larger than the other.

In our system, a route is a sequence of street segments leading from the origin to the destination. We consider every connection from one segment to another as an "intersection", even if there is only one next segment at the intersection. At any moment, the car will be on one of the segments of the route, approaching an intersection (unless an error occurs, which is handled as discussed below). The task of the Back Seat Driver is to say whatever is necessary to get the driver to go from the current segment, across the intersection, to the next segment of the route.

The items in the taxonomy of intersection types are called **acts**. We use an object oriented programming methodology, so for each act there is a corresponding "expert". The Back Seat Driver generates speech by consulting these experts. At any moment, there will be exactly one expert in charge of telling the driver what to do. To select this expert, the Back Seat Driver asks each expert in turn to decide whether it applies to the intersection. The experts are consulted in a fixed order, the most specific ones first. The first expert to claim responsibility is selected. This expert then has the responsibility of deciding what (if anything) to say.

- CONTINUE
- FORCED-TURN
- TURN-AROUND
- TURN
- FORK
- ENTER
- EXIT
- ONTO-ROTARY
- EXIT-ROTARY
- STOP

Figure 2: Act taxonomy

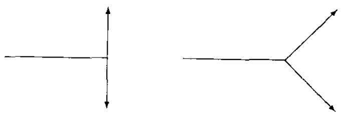

Figure 3: A "T" and a "Y" have the same topology

Describing actions

Each expert is able to generate text which describes the intersection. A description for an act must tell the driver two things: what to do and when (or where) to do it. "What to do" is expressed by a more or less constant verb phrase which depends upon the taxonomic classification, but may also depend upon specifics of the intersection. Thus a slight turn might be described by the verb "bear" where a sharper turn would be a "turn". The descriptions can be verbose or brief, and they can be expressed in past, present, or future tense. (We'll say why this flexibility is needed below.)

Saying "when"

Our study of natural instructions showed us that people almost never use distance as a cue for when to act. This is in sharp contrast to the textual directions provided by systems such as that of the Hertz rental company. Instead, people use two strategies. They wait until the driver is close to the intersection before saying anything, and/or they use a great variety of landmarks – including traffic lights, stop signs, other signs, buildings, road features, and the positions of other moving objects (e.g. "Follow that car."). The Back Seat Driver adopts both of these strategies.

Speech is especially useful as a cue for timing because speech is a temporal event, with a clear beginning and ending time. You know when someone begins to speak and when they finish. Someone peering at a map displayed on a CRT may have trouble distinguishing two adjacent streets, but there is no mistaking the word "now". Using time as a cue minimizes the workload on the driver, because the navigator absorbs the burden of remembering when to act. It also demands that the navigator have an accurate idea of where the car is. Our system demands positional accuracy of no greater than 10 meters for successful operation.

The Back Seat Driver's use of landmarks is unique in vehicle navigation systems. Our database began as a DIME file, but we extended it to include traffic lights, stop signs, road features (such as overpasses, bridges, and tunnels), distinctive signs, and the location of gas stations. Most of these are represented as attributes of the segments in the map database. To select a landmark for an intersection, the Back Seat Driver looks backwards from the intersection for the closest landmark which is also unique – that is, it prefers to say "take the first right after the underpass" rather than "take a right at the second set of lights". We think this makes the landmark easier to remember.

The Back Seat Driver does not speak at every single intersection. In the great majority of cases, it is perfectly obvious to the driver what to do (namely, to continue on forward). The action experts are also capable of deciding when the action at the intersection should be obvious to the driver. At present, the only action that is *ever* treated as obvious is CONTINUE. It is usually obvious to continue across an intersection, but we have found that what is obvious to one driver may not be so to another. Some people, for instance, are not comfortable driving across a major intersection unless they are instructed to do so. The expert can be somewhat customized so that its judgment of "obviousness" will correspond to that of the driver. If the action at the next intersection is obvious, the Back Seat Driver says nothing about it, and looks ahead for action at the next intersection, until it finds one that is *not* obvious.

The Back Seat Driver gives instructions just prior to the action. It also gives instructions further in advance, if time permits. This is especially useful when the instructions are complicated, as they are at some intersections. It is also able to give instructions "on demand". We call this the "what now" button. Drivers use this button for two reasons. Sometimes they are unsure whether they have come to the place where they are supposed to act, so they press the button to find out. At other times, they reach an intersection where the Back Seat Driver says nothing, because it believes the action is obvious, but it is not obvious to the driver. When the driver hits the "what now" button, the expert for the upcoming intersection describes it, even if it is considered to be obvious.

Talking about past and future

An advantage of language over pictures or gestures is that it can express events in the past or future. This advantage is well appreciated by readers of fiction, but may not yet be appreciated by designers of navigation systems. A navigation system should be able to talk about the past and future of the route, not just the present.

Drivers often need advance notice to prepare for an action. An example is what we call **lane advice**, which tells the driver to get into, or stay out of, a given lane. Lane advice is common in natural directions, and is one of the most appreciated features of the Back Seat Driver.

One reason for talking about the past is to describe mistakes. Drivers do not always follow the route the Back Seat Driver intends, either because of a mistake by the driver, the program, or external circumstances. When a mistake occurs, the Back Seat Driver finds a new route from the current location to the destination, while the driver is still moving. It also describes the mistake, saying something like "Oops, I meant for you to go straight." We think it is important that the system tell the user that there has been a mistake (without casting any blame on the user!) so that the user will come to better understand the system's style of instruction giving, and so that the user will remain confident in the system's understanding of the route. Talking about past and future actions is important in navigation. Speech seems to be the easiest way of doing this.

Example

As an example, here's a sample of the description of the left turn from Fulkerson Street to Main Street in Kendall Square, Cambridge.

> Get in the left lane because you're going to take a left at the next set of lights. It's a complicated intersection because there are two streets on the left. You want the sharper of the two. It's also the better of them. After the turn, get into the right lane.

This description was generated by the TURN expert in verbose form. It begins with some lane advice, then specifies the next action and provides a landmark for the place. The turn is described, and the proper street is described by two independent cues, one geometric, and one qualitative. Finally, the text provides a second piece of advice for after the turn.

Summary

The speech interface of the Back Seat Driver provides instructions without requiring the driver to look away from the road. Using speech permits us to talk about the past and the future as well as the present, and to give more detailed descriptions of the act than are possible with maps. Furthermore, it allows us to specify timing with great precision. But speech is not without its problems. The next section will discuss them, and the steps we have made to overcome them.

Liabilities of Speech

The advantages of a spoken language interface, as described above, do not come without cost. First, there are problems common to any natural language interface: while it is not terribly difficult to make a rudimentary interface, language generation requires substantial programming effort to be fluent and natural. Language is complicated, and people have literally a lifetime of experience with it, and are sensitive to fine nuances. On the other hand, having made this effort, we can exploit these nuances to to convey extra information.

A second problem is that a natural language interface is only useful to those who speak the language. In our experience, only a few non-native speakers have been able to understand the directions. Map displays have conventions of their own, but are more universal than natural language. We have also noticed that some driving terms used in the Boston area (e.g. "rotary") are not in the dialect of other English speakers. In our view, universality is not a prime concern. We believe that systems should be custom fit to the idiosyncrasies of their owners. The Back Seat Driver in your car should speak to you in the language and terms that are best for you as an individual, not you as a generic human.

The remainder of this section discusses problems specific to *spoken* natural language generation.

Speech takes time

As we said above, speech is inherently temporal. We take advantage of this when we use speech as a timing cue, but it also can be a difficulty. A real time spoken navigation system must plan its speech to ensure that it has enough time to say what it needs to say. If little time remains, it must say less (or speak more quickly), or ask the driver to slow down. We handle this problem by tracking the vehicle's position and velocity, and by modeling the time required to speak. The Back Seat Driver begins its speech at a time chosen to be early enough to allow the driver to hear the entire message, understand it, and react to it, before the point where action must be taken. The model of reaction time includes a constant for the driver's comprehension and a variable time which depends on the speed of the car, according to the maximum comfortable braking deceleration.

The temporal nature of speech also requires that the Back Seat Driver sometime combine instructions into a single utterance. When uttering an instruction, the Back Seat Driver looks ahead for the next instruction. If it determines that the time between the end of the execution of the current instruction and the beginning of the next is too short to allow it to speak the next instruction, it combines that text into the current one.

The Back Seat Driver does more than just give directions. Among other things, it also reads electronic mail messages from our office, gives weather reports, and makes comments about the route and road. Because speech takes time, and because a spoken utterance is only useful if completely spoken, the Back Seat Driver must carefully allocate the right to speak among potential tasks. It is undesirable for one task's speech to interrupt another's.

Speech can be misunderstood

A liability of speech, and synthetic speech in particular, is that speech can be misunderstood. This is particularly a problem with street names, because there are constraints that can help a driver correct a partially misunderstood name. A driver hearing an utterance that sounds like "Tarn reft" can guess that it is a corrupt form of "Turn left", but nothing can help the driver know what was intended by "Tarn Street". Directions should not use street names, because street name signs may be hard to see, misaligned, or simply missing. The importance of this first became apparent when we observed one driver who consistently misunderstood names, but also did not realize that he had misunderstood. Furthermore, the strength of his faith in the name was so strong that he drove straight through intersections, despite being told to "take the next left". This is probably the right thing to do with human instructions, where names are usually correctly understood, but street counts (e.g. "the third right") are imprecise or simply wrong. Our directions are phrased to minimize the use of street names in instructions. A typical text is: "Take the second left. It's Franklin Street."

Speech is transient

Information presented by speech does not persist, except in short term memory. We have already mentioned this as a reason why instructions should be given as late as possible. Another consequence of the transience of speech is that the system must be able to repeat itself at anytime, since the driver may not always be able to hear the speech. Repetition in turn poses a challenge.

since, unlike a program which reads the newspaper aloud, a literal repetition may not be appropriate, since the situation changes over time. For instance, if asked to repeat "Take the third left", the system may instead say "Take the second left" if the car has crossed an intersection. The consequence for the implementation is that the system retains not its previous words, but rather the previous reason for speaking. When asked to repeat, it invokes the same function that produced the last utterance.

A second problem with the ephemeral quality of speech is that the driver has no evidence of the program's existence except when it is speaking. We consider it very important that the driver have continued confidence that the program is running correctly, is aware of the driver's position and progress, and is "seeing" the world in the same way the driver does. We have devoted substantial effort to maintaining the illusion of **co-presence**.

In the introduction to this section, we said that the nuances of language could be used to convey much information. Co-presence is an idea communicated more by nuance than by explicit statement. (People would laugh if the system said "I'm right here with you." It sounds like something a therapist would say.) One way we indicate co-presence through nuance is by using **deictic** pronouns. Deictics are words that "point" at something. In English, we have four deictic pronouns: "this", "that", "these", and "those". The first two are singular, the second plural. The difference between "this" and "that" (and "these" and "those") is that "this" refers to something close. We use this in referring to landmarks. When the landmark is close, we use the proximal form (e.g. "these lights") ; when distant, we use a brief noun phrase (e.g. "the next set of lights"). This is important. When a driver is stopped 30 meters back from a stop light, it may be literally true to say "turn left at the next set of lights", but it will confuse the driver.

A second means of conveying co-presence is to acknowledge the driver's actions. After the driver carries out an instruction the system briefly acknowledges the act if there is time, and if the act was not so simple (e.g. continuing straight) as to need no acknowledgment. This acknowledgment is a short phrase like "Okay". Some drivers dislike acknowledgments, so they can be disabled, but most find the confirmation comforting. The timing of the acknowledgment does much to confirm the driver's sense that the program really knows where the car is. Another source of acknowledgment is the use of **cue words** in the instructions. It will often by the case that the route calls for the driver to do the same thing twice (e.g. make two left turns). The speech synthesizer we use has very consistent pronunciation, and drivers sometimes get the impression that the system is repeating itself because it is in error (like a record skipping). The acknowledgments help to dispel this, but we also cause the text to include cue words such as "another". These indicate that the system is aware of its earlier speech and the driver's previous actions.

Yet another means of conveying co-presence is to make occasional remarks about the road and the route. These remarks indicate that the program is correctly oriented. As an example, when the road makes a sweeping bend to one side, the program speaks of this as if it were an instruction ("Follow the road as it bends to the right.") even though the driver has no choice in what to do. The program also warns the driver about potentially hazardous situations, such the road changing from one-way to two-way, or a decrease in the number of lanes. As with acknowledgments, these warnings can be disabled if the driver dislikes them. Other remarks have less to do with the route. We justify these by the maxims of cooperative conversations formulated by philosopher H. P. Grice[2]. His maxim of QUANTITY (part 1) says: "Make your contribution as informative as is required." Grice explains that one can convey information by appearing to flout the maxim. In this case, a driver can reason as follow: "The program, like all cooperative agencies, obeys the maxim of quantity. Therefore, it is had something important to say, it would say it. The program said nothing of great significance, therefore there is nothing urgently requiring my attention. So everything is well." At present, our "Gricean" utterances are trivial observations about the weather, but we are re-designing them to convey useful information about the city.

Summary

A speech interface for giving driving instructions has several advantages over a graphics interface. There are problems with natural language interfaces in general, and speech in particular, but they can all be overcome. The result is an excellent aid for navigation.

Acknowledgments

The authors wish to gratefully acknowledge the support of NEC Home Electronics, Ltd.

References

[1] James Raymond Davis. *Back Seat Driver: voice assisted automobile navigation.* PhD thesis, Massachusetts Institute of Technology, September 1989.

[2] H. P. Grice. Logic and conversation. In Cole and Morgan, editors, *Syntax and Semantics: Speech Acts*, volume 3, pages 41–58. Academic Press, 1975.

[3] Osamu Ono, Hidemi Ooe, and Masahiro Sakamoto. CD-ROM Assisted Navigation System. In *Digest of Technical Papers*, pages 118–119. IEEE ICCE, 1988.

[4] Christopher M. Schmandt and James R. Davis. Synthetic speech for real time direction giving. *IEEE Transaactions on Consumer Electronics*, page (to appear), 1989.

[5] Lynn A. Streeter and Diane Vitello. A profile of drivers' map reading abilities. *Human Factors*, 28:223–239, 1986.

[6] Lynn A. Streeter, Diane Vitello, and Susan A. Wonsiewicz. How to tell people where to go: comparing navigational aids. *International Journal of Man/Machine Systems*, 22(5):549–562, May 1985.