Investigating a Two-Way-Audio Query-Response Command Interface with Navigation Data Extraction for Driver Assistance

Ronald P. Loui, Vishakha Kharpate, Mitali Deshpande, Fatmah Alanazi
Department of Computer Science
University of Illinois
Springfield, IL

***Abstract* — Navigation systems for drivers tend to vocalize what they want, when they want. Drivers who need to clarify instructions are often required to read directions in text form, which is a safety risk during real-time control.**

Instead, we investigate how a two-way-audio command interface might work. Data is easily extracted from text directions so that questions can be answered without knowledge of the GPS location, as long as the system is aware of the most recently vocalized directional cue (and elapsed time since last utterance). Functions envisioned are as simple as scrolling forward or back through the directions, spelling names or pronouncing differently, repeating the road name or directional change, and saying what is next on command. The queries require modest data extraction with potentially great improvement in usability. Queries may further refer to time and the results of some intermittent internet search, but do not attempt to interpret 2d map data.

This project is ongoing, and the contribution includes a novel design architecture that puts a second AI assistant in the co-pilot's seat. This design provides domain-specific memory and vocalization assistance on top of the navigational assistance already familiar to the driver: voice AI (vAI) for improved voice UI (vUI).

It aims more for a smart TiVo than co-pilot Siri. When navigation assistance and two-way audio dialogue are provided by the same vendor, the audio recognition and current location errors can be significantly reduced, but the current design is intended as a third-party intervention, adding assistance where the original product is lacking and possibly creating pressure for their products to improve.

I. Introduction

Most would agree that text comprehension has no place in the driver's seat. Even as sophisticated sensor-based semi-autonomous driving modes find their way into consumer vehicles, real-time command and control in many situations still depends on near-constant driver attention.

GPS-based navigation guidance is popular, whether provided in-dash or with maps and apps on phones. Voice-based cues for turns, exits, and merges are popular and successful. However, with a small amount of user-interface design, the usability could be much improved.

We have investigated two-way audio with query-response dialogue, as an innovation and improvement on the dominant one-way audio interface paradigm for voice-guidance. The key idea is that the text directions can be mined for simple answers to specific queries. The crucial added functionality is the driver's ability to query the system, rather than wait to be told by the system what to do. Thus, we supplement the vocalization of prompts with a query-response, or question-and-answer-extraction system.

Apple's Siri, Amazon's Alexa, and OK Google are leading voice-based, two-way audio command interfaces. These consumer products seek to be general purpose assistants that leverage much of their power from internet search and some cyber-physical system integration (e.g., with phone or audio playback subsystems). Our approach is the opposite: only those functions that are most useful to the driver, based on the data contained in the driving directions text, are within our automated assistant's scope. The design is for the automation of just a few co-pilot functions in a noisy, error-prone, high-risk environment, and the depth of AI ambition is limited, with these important task-specific issues in mind.

Several automobile companies have been deploying minor voice-controlled command functions for their in-dash "infotainment" systems. So far, these command functions pertain to menu selection, in lieu of physical touch-screen activity (e.g., Honda, Ford [1,2], which include few navigation direction voice commands). The idea here is to address the text directions, not the dashboard control menu selections. We would prefer that the companies building the navigation assistance software do two-way assistance integration, with their navigation systems as we propose, so that location and voice-recognition error can be minimized. However, it is an interesting design problem to deploy the two-way audio assistance as a third-party phone app (or laptop program) that is separate from the in-dash navigation and separate from a dedicated GPS navigation device (e.g., a Garmin unit). This is in part to support legacy vehicles. But it is also to bypass the entrenched preferences, or slow R & D, of those consumer-driven industries. It provides an important intervention for designers who desire to push navigation vendors to improve their user interface, as a competitive, added-value platform.

978-1-5090-4767-3/17/$31.00 ©2017 IEEE

The envisioned voice AI aimed at improving voice UI is itself an interesting platform, which might be useful in other settings beyond automotive assistance. This introduces the "third eye" or "third AI" into the room, which is mainly aimed at supporting clarification subdialogues. Since the primary functions we consider are rewind and re-vocalize, pause, and clarify (e.g., spell what was said), the objective is more like a commandable audio TiVo than Siri or Alexa.

II. ADDING CO-PILOT E-VISHAKHA

The first set of commands illustrates the huge usability gap that exists in the current one-way, mono-pragmatic monologue on which current products are based. Rather than be a passive, commanded driver, we change the paradigm and enable a commanding driver to utter:

REPEAT!
REPEAT SLOWER!
REPEAT LOUDER!
SAY DIFFERENTLY!
SPELL THAT!
WHAT'S NEXT?
WHAT'S AFTER THAT?
WHAT WAS BEFORE THAT?
WHAT ROAD IS THIS?
WHAT ROADS ARE NEXT?
LEFT OR RIGHT NEXT?
DETAILS ON!
DETAILS OFF!
ENOUGH!

This basic command set includes both imperatives and interrogatives. It implicates AI at the sound level, the speech act performative level, the dialogue level, and the meta-dialogical sub-dialogue (repair) level. Design requires locutions that are unambiguous and easily remembered. For example, "FORWARD!" and "BACK!" might confuse direction of motion with progress through the list of directions. There might also be existing UI kinds of forwards and backwards, e.g., "FORWARD" as "SCROLL FORWARD!" which is further ambiguous between text and screen.

Of course, to improve sensitivity, the system should require that the driver precede each command with a unique name. We arbitrarily use "VISHAKHA" as that name. As with Siri, the system's name is chosen to be phonetically distinct and conversationally improbable. One problem with in-car voice recognition is that mic sensitivity often varies with background db level, so parts of side conversations are intermittently recognized, and this can cause false-commanding. "Thus, REPEAT!" must be prefixed with "VISHAKHA, REPEAT!" or "VISHAKHA, PLEASE REPEAT!"

All of these commands were tested using a Samsung Galaxy S3 with Android and Google voice recognition (keyboard option), with 100% recognition of the commands. None were misspelled, and no extraneous words or syllables were recorded. "Vishakha" was not recognized as an English language phrase consistently, which is desirable because it is a delimiter (it was misrecognized as "The Shaka", "The Shocker", "This Chaka", and "Cachaca" in addition to "Vishakha", where the latter was likely based on text mining the user's email). Presumably this attention word can be learned by the speech to text program. Of course, the commands and system attention phrase can be revised without changing the basic proposal, so long as disambiguation, brevity, and ease of recall are kept in mind.

There is a natural trade off between providing many alternate forms of giving commands, and creating a lexicon of utterances that is too large, hence, not as easily disambiguated. A "phonetic Hamming distance" based on fricatives (which vary less than vowels between different speakers), stress rhythm, and syllable length is desirable for any command set. A large command set does not always provide easier recall, not even when a single member of an equivalence class suffices. Our preference is to have "HELP!", "MENU!", and "LIST COMMANDS!" generate an auditory review of the options, rather than to add such things as "TELL ME WHAT I CAN SAY!" to the command set.

The next set of commands requires minor estimation, calculation and search, not just attention to the sequence of directions. Where [X] occurs in the command, the voice to text requirement is greater, but the target text is constrained by the words and word pairs occurring in the text of the driving directions.

HOW MUCH LONGER?
WHERE WILL WE BE IN AN HOUR?
HOW LONG UNTIL THE NEXT EXIT?
HOW LONG SINCE [X]?
HOW FAR TO [X]? (internet search)
WHAT IS THE SPEED LIMIT HERE? (internet search)
IS THIS A TOLL ROAD? (internet search)
WHAT IS THE TOLL? (internet search)

The system may require "ESTIMATE CURRENT ROAD SPEED" and may simply respond "UNKNOWN", "DON'T RECOGNIZE [X]", or "[X] IS NOT ON THIS ROAD". But if the driver accepts incomplete or imprecise answers as part of the pragmatics, the commands are still implementable and potentially useful. The system may also ask the driver to verify what (what road) the current position is, because of missed vocalizations (see below).

Directed internet search to determine the distance between two points is easy. Finding the speed limit on a segment of road will not always be possible unambiguously. Since we do not assume access to GPS location, when deployed as a third-party intervention, some of the queries are answered by taking the time since last uttered instruction by the unprompted navigation, and an estimate of the road speed (based on the type of road or the user's input).

The key idea here is to generate answers from text (possibly text resulting from specific search), not to attempt interpretation of 2D mapping data.

III. DATA EXTRACTION FROM TEXT

Many navigation directions appear in overview form, with details for each segment. For example, google's driving directions between UIS and UND have the overview:

```
Get on I-55 N in Woodside Township from W Lake Shore Dr and Stevenson Dr

8 min (3.3 mi)

Continue on I-55 N. Take I-39 N/US-51 N, I-90 W and I-94 W to 34th St S in Moorhead. Take exit 2 from I-94 W

10 h 40 min (753 mi)

Continue on 34th St S. Take US-75 N and 15th Ave N to Elm St N in Fargo

14 min (6.4 mi)
```

with the following details for the first segment:

```
Head northeast toward Eliza Farnham Dr

495 ft

Turn left onto Eliza Farnham Dr

0.1 mi

Turn right toward Edgar Lee Masters Dr

0.1 mi

Turn right onto Edgar Lee Masters Dr

0.2 mi

Turn left onto University Dr

0.4 mi

Continue onto W Lake Shore Dr

1.3 mi

Turn right onto Stevens Dr

0.7 mi

Turn left onto the Interstate 55 N/Interstate 72 E/Route 36 E ramp

0.3 mi
```

If we were to extract data from the overview, by recognizing and dropping the key and stop words, we might mine the text fragments:

< I55, N, Woodside Township, W Lake Shore Dr, Stevenson Dr , 8 min, 3.3 mi>

for the first part of the overview.

Without actually knowing which of the four proper names refers to the current road, the response to “WHAT ROAD IS THIS?” would be “I55. WOODSIDE TOWNSHIP. W LAKE SHORE DRIVE. STEVENSON DRIVE.” Rather than take the risk that the prepositional indicators can be parsed correctly for sequential and contextual information, we assume that this answer suffices to an intelligent operator who may already be familiar with the text directions. The driver can always ask for a “REPEAT!” of the last instruction if this answer were unintelligible.

Fortunately, turn-by-turn instruction takes place on the details, so there is no ambiguity in any of the instructions, except the last. “INTERSTATE 55 N. INTERSTATE 72 E. ROUTE 36 E.” would be the response. It actually names a ramp, not a road segment, but to the intelligent human receiver, given the limited pragmatic intent, this makes a reasonable trade off between risk of error and likelihood of usefulness.

The referential objects extracted from the text are only the proper names, the lateral directions, LEFT and RIGHT, and the keyword EXIT. The situation semantics makes the current road the default referent and point of reference.

To answer the request to “SPELL THAT!”, there may be a similar ambiguity. For “E D G A R SPACE L E E SPACE M A S T E R S SPACE D R” we would want to spell each word in sequence until “ENOUGH!” is commanded. For the ambiguous case, the response would be “I FIVE FIVE N. I SEVEN TWO E.” A few “expert system” rules would be helpful, such as spelling “I” instead of “I N T E R S T A T E” and converting “Dr” to “DRIVE”. We would also add rules to change “0.2 mi” into “SHORT DISTANCE” both for comprehensibility and lucidity.

IV. SPEECH 2 TEXT AND TEXT 2 SPEECH

Implementation of this design currently focuses on two issues: text extraction with stop words ignored, and special rules for responses, such as saying “LEFT OR RIGHT IS UNKNOWN”.

Recognizing the command set in a noisy environment with varying levels of road noise, accents, listening angles, and speaking volumes is a challenge. When a phone app is listening to both the driver and the machine-generated navigation, as well as background chatter, there will be vocalizations missed and others mistaken.

Using existing third party speech-to-text in actual car-in-traffic experiments, we discovered that human commands are easy to recognize, even with nuisance conversation, but the navigation instructions had a surprisingly high rate of omission as well as mis-recognition. The latter can be repairable with matching of expected text with recognized text, or it can be unrepairable and simply garbled.

For example, with one phone’s microphone next to another phone’s speaker, in normal traffic, using an iPhone to generate directions and driver navigation, and an Android/Google voice recognition API:

Chatham Road.

You should reach your destination by 2:14.

Turn right onto Wabash Avenue. Continue for two miles.

Was recognized perfectly. But

Furniture Good 55 Business Loop South I

was supposed to be “Turn right onto I 55 Business South” (the word “Loop” is not in the text version of the directions).

Time Left

was supposed to be “Turn left onto Hazel Dell Road” and

In 600 feet turn right onto

was supposed to name the street “Octavus Via” which was clear to the human ear, but not to the speech recognition program. These names would likely be recognized if we had emailed the instructions so gmail could be mined for proper nouns. It is not clear whether the design can assume novel proper noun recognition with such priming; better to design around the low-fidelity recognition of names. One could work with “Road” instead of “Hazel Dell Road”, but the chance of misalignment is significant enough to require a workaround, such as **eVishakha** qualifying a response with “THE LAST INSTRUCTION I HEARD WAS ...”. This admits to potential omission and reduces risk of overconfidence in the AI.

Finally,

Turn right on Division Drive

should have been “Turn right on Lejeune Drive”.

On the return pass,

Head East on Lejune Drive toward Octopus Via

captures most of the name, dropping an “e”, but misses “Octavus”,

Turn Left onto

misses the name, “Octavus Via”, completely. Then

In a quarter mile, turn left onto Lincolnshire Boulevard

correct but for a small issue matching “Boulevard” to “Blvd”, then

and turn right on I-55 North

with a slight mismatch to the instructions over stop words, “on” and “onto”.

2 Stanford Avenue on E Stanford Avenue

is written as “turn left onto E Stanford Ave”, but is still easily matched as the expected next major road name.

Continue for one and a half miles

is vocalized twice, with no representation in the text instructions, in part because Stanford Avenue and Wabash Avenue are the same road, and “Wabash Ave” is in the text. This demonstrates a frustrating reality of voice to text: a name recognized earlier, vocalized in exactly the same way with different context and noise, may not be recognized the next time it is uttered.

In a quarter mile, turn left onto South Park Avenue

In 1000 feet use the left two lanes to turn left onto Chatham Road.

The last instruction, “Chatham Road” permitted recovery of alignment with instructions after taking an alternate route that deviated from the exact instructions. This shows that the ability to re-align after omission or mis-recognition is similar to the ability to re-align after minor deviation from route.

iPhone navigation instructions tend to differ from text instructions more than other mapping software (it tends to give audio commentary that is not in the written text), but the potential for misalignment is real with any navigation app. One might deliberately target the **eVishakha** application to navigation assistants where vocalization and text directions are close. For the iPhone speech being recognized by Samsung/Android/Google, we estimate text over-vocalization at about 10%, which can be safely ignored, text under-vocalization to be about 10%, which requires skipping ahead to realign, and omission or name recognition failure at about 15%.

Note that template-based text generation and driver recognition of text-to-speech are not considered to be problems. We prefer to deploy a laptop-based program for better control of the voice-to-text recognition, including possibly changing the parameters to favor automated voices over human voices.

Additional experiments were performed using various text-to-speech (Text Aloud, Yunus, iPhone, Balabolka) and speech-to-text programs (Speech Note, Dragon by Nuance, Voice2Text by Christian Neubauer) on actual navigation instructions with varying background noise conditions. The following table summarizes these experiments.

Clearly the keywords “TURN” and other common words are easily recognized, but street names are challenging without some additional processing to make matches. Distances were also reliably recognized. The compass directions were challenging when compound, e.g., “Northeast”, or when they occurred in the first command (i.e., at vocalization “cold start”).

	recognized	recognizable	garbled	omitted	% raw	% w/AI
TURN	69				100.00	100.00
LEFT/RIGHT	104	1	1	1	97.20	98.13
E/W/N/S/NE/SW...	27		2	4	81.82	81.82
NAME	106	42	9	4	65.84	91.93
distance	74		2	1	96.10	96.10

V. CONCLUSION

Since the system relies on the last utterance (and time since last utterance) to determine sequential (and spatial) position, a missed vocalization from the GPS navigation system can produce error. Of course, with any GPS-based, static map-based, and audio-based system, error is potentially present. Another source of error is when the vocalizations do not actually match what is in the text directions. That is precisely why it is important to design the system so that robustness to misinterpretation and/or mis-recognition is constantly a concern. We rely mainly on the recognition of turns and matching main road names, even when recognized imperfectly. It is unfortunate that major road names are not repeated by the navigation system, because they are the ones that permit best, and longest re-alignment after losing alignment.

A similar problem is when the system re-routes and produces new driving directions, at least for one segment's detail. For minor deviations, the re-routing is roughly the same as a missed speech-to-text match. For major deviations, where re-alignment takes some time, the driver receives no useful assistance for that time.

The primary observation from this preliminary work is that useful data can indeed be extracted from text and matched to some extremely useful, albeit unspectacular commands. Re-vocalization, preview of instructions, and memory assistance provide improved user interface in this important and ubiquitous form of driver assistance. The assistant, and the driver both need to be aware of the potential for error. If there were better ways to bias the voice recognition in favor of the names in the text directions, that would be helpful. Ironically, it appears that the voice recognition is tuned to human voice frequencies, so ironically, automated text to voice is harder to recognize than human speech.

Beyond the design balance between artificial intelligence (AI) and user interface (UI), we introduce the idea of a second co-pilot, or a second AI, whose purpose is to improve the UI. In a different application, such an "AI-4-UI" or "vAI-for-vUI" (voice AI for voice UI) might translate from GPS navigation instructions in one language, such as English, into another, such as Spanish. Here, most of what we have asked the system do to is to act as a glorified TiVo, but even that low tech ambition may have a great impact on safety given the important co-pilot position it occupies.

With Siri and Alexa, we see a wide range of behaviors attempted for voice-commanded assistants. Meanwhile we have seen little imagination or human accommodation in the emerging systems specifically designed for driver assistance. How and when information can and should be presented to the driver may require dialogue, not just well-programmed monologue. There is room for design in the space, even if it is just a way to encourage user interface improvement among the existing navigation system vendors.

ACKNOWLEDGMENT

The authors thank the Graduate Research Seminar for their t2s2t measurements, with teams led by Benarjee Chitekala, Mahender Bhonagiri, Adithya Ampati, and Venkateswarlu Namburu.

UIS Tech Brainstorming Club gave design ideas, esp. Ehtesham Syed. Mahathi Karnamadakala had many insights from his work at Google Maps that were too ambitious to include here.

REFERENCES

[1] Honda Voice Command Index 2017 downloaded from techinfo.honda.com.

[2] Shutko, John, Ken Mayer, Eero Laansoo, and Louis Tijerina. *Driver workload effects of cell phone, music player, and text messaging tasks with the Ford SYNC voice interface versus handheld visual-manual interfaces*. No. 2009-01-0786. SAE Technical Paper, 2009.

[3] Weng, Fuliang, Pongtep Angkititrakul, Elizabeth E. Shriberg, Larry Heck, Stanley Peters, and John HL Hansen. "Conversational In-Vehicle Dialog Systems: The past, present, and future." *IEEE Signal Processing Magazine* 33, no. 6 (2016): 49-60.

[4] Ramaswamy, Ashok B., and Randall T. Brunts. "Mapless GPS navigation system in vehicle entertainment system." U.S. Patent 5,627,547, issued May 6, 1997.

[5] Trovato, Karen I., and Daniel L. Pelletier. "Portable system for providing voice driving directions." U.S. Patent 5,835,881, issued November 10, 1998.

[6] Junqua, Jean-claude. "Method for natural dialog interface to car devices." U.S. Patent 6,598,018, issued July 22, 2003.

[7] Weber, Dean. "Object interactive user interface using speech recognition and natural language processing." U.S. Patent 6,434,524, issued August 13, 2002.

[8] Chu-Carroll, Jennifer. "Form-based reasoning for mixed-initiative dialogue management in information-query systems." In *EUROSPEECH*. 1999.

[9] Berry, Michael W. "Survey of text mining." *Computing Reviews* 45, no. 9 (2004): 548.

[10] Zhang, Xiao, Prasenjit Mitra, Sen Xu, Anuj R. Jaiswal, Alexander Klippel, and Alan M. MacEachren. "Extracting Route Directions from Web Pages." In *WebDB*. 2009.

[11] Zhang, Xiao, Prasenjit Mitra, Alexander Klippel, and Alan MacEachren. "Automatic extraction of destinations, origins and route parts from human generated route directions." In *International Conference on Geographic Information Science*, pp. 279-294. Springer Berlin Heidelberg, 2010.

[12] Moncla, Ludovic, Mauro Gaio, and Sébastien Mustiere. "Automatic itinerary reconstruction from texts." In *International Conference on Geographic Information Science*, pp. 253-267. Springer International Publishing, 2014.

[13] Zhang, Xiao. "Route extraction, road name disambiguation and efficient spatial query processing under location constraints." PhD diss., The Pennsylvania State University, 2012.

[14] Leijon, Arne, Gustav Eje Henter, and Martin Dahlquist. "Bayesian analysis of phoneme confusion matrices." *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 24, no. 3 (2016): 469-482.

[15] Fosler-Lussier, Eric, Ingunn Amdal, and Hong-Kwang Jeff Kuo. "A framework for predicting speech recognition errors." *Speech Communication* 46, no. 2 (2005): 153-170.

[16] Pisoni, David B., Howard C. Nusbaum, Paul A. Luce, and Louisa M. Slowiaczek. "Speech perception, word recognition and the structure of the lexicon." *Speech communication* 4, no. 1 (1985): 75-95.

[17] Litman, Diane J., and James F. Allen. "A plan recognition model for subdialogues in conversations." *Cognitive science* 11, no. 2 (1987): 163-200.