Speech and sound for in-car infotainment systems

Ivan Tashev, Michael L. Seltzer, Yun-Cheng Ju
Microsoft Research, One Microsoft Way, Redmond, WA 98052, USA
{ivantash, mseltzer, yuncj}@microsoft.com

ABSTRACT

In a hands-busy and eyes-busy activity such as driving, spoken language technology is an important component of the multimodal human-machine interface (HMI) of an in-car infotainment system. Adding speech to the HMI introduces two distinct challenges: accurately acquiring the user's speech in a noisy car environment, and creating a spoken dialog system that does not require the driver's full attention.

Categories and Subject Descriptors

H.5.1. [**Multimedia Information Systems**]: Audio input/output H.5.2. [**User Interfaces**]: Natural language, Voice I/O, Interaction styles, Prototyping.

General Terms

Algorithms, Design, Human Factors.

Keywords

Speech interfaces, in-car infotainment, multimodal UI.

1. INTRODUCTION

Voice-enabled dialog systems are among the most attractive features of in-car infotainment systems, since speech is the only wideband communication channel not significantly engaged during driving. Such systems first began to appear in high end cars, but recently found their way to mass production cars, such as Ford and Fiat. Currently they are in transition from being a cool gadget to an integral part of the modern automobile. Speech interfaces are being integrated into a growing number of applications, ranging from control of the radio and other equipment in the car, to the use of external devices such as mobile phones, portable media players, and navigation systems.

2. SOUND CAPTURE SYSTEM

Environmental noise is the major challenge for sound capture in cars. The in-car system typically has a speech enhancement block that attempts to remove environmental noise, followed by a chain of several encoder/decoders (Bluetooth, GSM, G711). Another challenge is to make the sound capture system suitable for both human-human communication and speech recognition (both in-car and server-based). To maximize the perceptual sound quality and speech recognition results end-to-end optimization of the system can be performed as described in [6]. The sound capture can be further improved by adding more microphones to form a microphone array [5].

3. NATURAL LANGUAGE INPUT

Most conventional in-car speech systems work with a fixed grammar, i.e. a set of commands that the driver has to remember. For improving the system usability, we propose using a more flexible system that replaces the grammar with a statistical language model (SLM). For added robustness to system and user errors, the recognizer output is then post-processed using information retrieval techniques commonly used in web search, such as TF-IDF. This combination of SLM-based speech recognition and search techniques enables the system can find the most relevant match for the user's request even when the input query is not exact. Using this approach, we have designed applications for selecting music from a media player [3][4], replying to text messages [2] and searching the car owner's manual.

Speech is a very useful and efficient input modality when selecting an item from a long list, e.g. 5K songs. However, the usability and efficiency of speech decreases if the list length is small, e.g. selecting the exact song from the top four hypotheses. In these situations, when the driver can see all four candidates by just glancing at the screen of the in-car system, touch or button may be a preferred input modality. A multimodal user interface that appropriately combines the strengths of speech, graphics, button, and touch is less distracting and more convenient than a user interface that relies exclusively on any one single modality.

4. CONCLUSIONS

A successful human-machine interface for in-car applications is one that allows drivers to perform non-essential infotainment tasks without adversely affecting driving performance. Along with touch, buttons, and graphical displays, speech is a key modality that enables the design of user interfaces that can improve usability and reduce distraction for drivers. More information and videos of our prototype can be found on our project web page [1].

5. REFERENCES

[1] Commute UX project at Microsoft Research http://research.microsoft.com/en-us/projects/CommuteUX/.

[2] Ju, Yun-Cheng; Paek, Tim. 2009. A Voice Search Approach to Replying to SMS Messages in Automobiles. Proceedings of Interspeech, Brighton, UK.

[3] Ju, Yun-Cheng; Seltzer, Michael; Tashev, Ivan. 2009. Improving Perceived Accuracy for In-Car Media Search. Proceedings of Interspeech, Brighton, UK.

[4] Song, Young-In; Wang, Ye-Yi; Ju, Yun-Cheng; Seltzer, Mike; Tashev, Ivan; Acero, Alex, 2009. Voice Search of Structured Media Data, in proceedings of Int. Conf. on Acoustics, Speech and Signal Processing, Taipei, Taiwan.

[5] Tashev, Ivan. *Sound Capture and Processing – practical approaches.* John Wiley & Sons, 2009.

[6] Tashev, Ivan; Lovitt, Andrew; Acero, Alex. 2009. Unified Framework for Single Channel Speech Enhancement. Proceedings of IEEE Pacific Rim Conference on Computers, *Communications, and Signal Processing. Victoria, Canada.*

Copyright held by authors and Automotive UI '09, September 21-22, 2009, Essen, Germany. Adjunct Proceedings, http://auto-ui.org.