Acoustics aspects of in-vehicle spoken dialogue

Klaus Linhard, Paul Heisterkamp
DaimlerChrysler AG, Research and Technology, Dialog Systems, Ulm, Germany; Email: {Klaus.Linhard|Paul.Heisterkamp}@DaimlerChrysler.com

Introduction

By making speech dialogue in the car easier and better to use, we aim at reducing possible driver distraction, and thus contribute to the DaimlerChrysler vision of ‘accident-free driving’. The car is known to be a 'hostile’ acoustic environment, not only because of the internal and external, stationary and in-stationary noise, but also because of the many hard reflecting surfaces. The market success of in-vehicle spoken dialogue systems shows that many of these problems are handled quite well today. On the other hand, the car is also a 'known' acoustic environment. We discuss an example of how this can be exploited for better performance of dialogue systems, and how knowledge from the dialogue side can be made use of in acoustics: The intelligent handling of barge-in, in particular in the presence of in-stationary noise. In these conditions, a purely energy-based barge-in detection often leads to a very annoying and confusing dialogue behavior. To reduce false alarms, we make use of speech recognition to identify barge-in utterances and to differentiate them from so-call back-channel utterances. We make use of dialogue knowledge to react appropriately to the respective types of user utterances, and we are currently investigating ways to handle speech synthesis such that the reactions to the barge-in are more human-like and thus less irritating than the current practice of having the system voice cut in mid-word. We conclude by giving an outlook on further work that couples acoustics, recognition, dialogue and speech output even closer for a better performance of the overall in-vehicle system.

Motivation

In vehicles today, many information, communication and entertainment systems are available. To make good use of these systems, the driver has to operate them. Operation of such systems, however, is only a secondary task for the driver: the primary task is to drive safely, with hands on the wheel and eyes on the road. This makes speech dialogue a natural choice as a channel for the operation of theses systems, as the acoustic channel not occupied by the primary task. Also, a well-working speech dialogue system enables a much more convenient handling of complex systems than screen-and-knob operation, as speech is natural and allows for direct access to specific functions, in particular if the dialogue management helps to overcome the limitations of vocabulary size and understanding capabilities. Speech, therefore, is a way to minimize potentially dangerous driver distraction. This is why a car manufacturer like DaimlerChrysler has been active in research in speech, dialogue and acoustics for quite a number of years.

Current in-car speech systems

In the S-Class car of 1996, Mercedes-Benz introduced the first generation of Linguatronic. Linguatronic is the brand name used in Europe of a speech dialogue system that allows completely hands-free operation of the car’s mobile phone, including number dialing (with connected digit dialog), number storing, user-defined telephone directory entry name, name dialing, and directory editing. Linguatronic I has a vocabulary of about 30 speaker-independent words (digits and control words). The second version (on the market since 1999) has a vocabulary of about 300 words, and, in addition, allows for operation of comfort electronics (radio, CD-player/changer, air condition etc). The system is now available for German, US English, UK English, and a number of other languages (cf. [1]) Some manufacturers, like BMW and Audi, also use this technology, which is developed into a product and marketed by Temics SDS, a former subsidiary of DaimlerChrysler, whereas others, notably Jaguar and Honda (Acura) rely on either self-developed systems or use those of other major speech technology companies like SpeechWorks or IBM. The market is expanding. Some more car makers will introduce speech systems in 2003-2004.

Acoustic processing

The acoustic pre-processing used in these systems are also used to enhance the speech quality for in-car mobile telephony. In Linguatronic, the same microphone and pre-processing is used for both the telephone and the speech dialogue system, not in the least for cost reasons. Cost for installation, processors and memory is also a major limiting factor for the use of microphone arrays.

Noise cancellation and echo compensation are important features. Echo compensation is particularly tricky. A passenger car has many windows. Windows are hard, reflecting surfaces, and they vibrate in themselves. Therefore, it is desirable to measure the echos for each model and make adjustments according to their delay and energy.

Stationary noise from the car’s engine and movement is handled very well by estimation during speech pauses. In-stationary noise originating from road bumps or large rain drops or other sources still poses a major challenge. Although these noises are only partly inside the frequency range that is important for speech dialogue systems, their existence prohibits the use of standard, energy-based barge-in as utilized in telephony systems. All too frequently some noise triggers the barge-in mechanism, leading to a very annoying system behavior. in. For speech recognition purposes, another important acoustic problems is off-talk: someone is speaking, but this is not intended for the system. In the following, we propose how a close co-operation of dialogue management and acoustics can help to overcome these problems.

Barge-in

Both in talking to other people and to speech dialogue systems, people do not wait for dialogue partners to finish speaking. Rather, they start speaking themselves as soon as prosodic, semantic and other cues indicate that it is not impolite. This is generally called ‘Talk-over’:

System: “So that would be for Monday, so for one night, right?”
Driver: “Right”

Talk-over can be un-interruptive, e.g. in acknowledging semantic items, assuring the communication channel is open and one is still listening. This behavior called ‘Backchannel’).

System: "So that would be for Monday, so for one night, right?"

Driver: "Yupp" "Right"

Backchannel is often realized using non-word sounds, written as 'hmm', 'eh-he' etc. These sounds, just like other human non-speech sound production, e.g. crying, laughing, sneezing etc. are called 'paralinguistic'.

"Barge-in" is a special case of Talk-over, where the interruptor 'grabs' the turn from the speaker, who normally stops speaking:

System: "So that would be for Monday, so for ..."

Driver: "No! For Tuesday!"

Backchannel and Barge-in allow humans to speed up communication. Speech dialogue systems normally need some form of barge-in mechanism to avoid user frustration and capture Talk-over. In standard telephony systems, the barge-in mechanism normally relies on some form of echo compensation for the system utterance. The speech recognizer's segmentation interrupts a system utterance whenever speech is detected. This is mostly dependent on energy levels in the frequency domain.

The main problem of the energy based approach to a barge-in mechanism is false alarms. This leads to very annoying system behavior that can be tolerated in telephony systems with the steep learning curve of speakers. In in-car systems, however, false alarms are not only due to a caller speaking at wrong times, but also to the in-stationary noises mentioned above, and to other people in the car speaking. This is why we pursue a close coupling of acoustics, recognition, understanding and dialogue to identify the source of the sound event that is a potential barge-in candidate, to classify whether it is noise, back-channel or real barge-in. (A somewhat similar approach is described in [2], but cf. also [3]).

First, we need classification and training of backchannel words and paralinguistic sounds, and, perhaps, some noises that have much energy in the speech frequency range. Using the speech recognizer and the semantic parsing, we can classify the talk-over utterance as backchannel, talk-over or barge in. The dialog manager then has to decide on how to react.

In the case of backchannel, it can increase the confirmation level of a semantic item that was addressed in the part of the system utterance immediately preceding the backchannel signal, provided the absolute timing of the synthesizer output is synchronized with the recognizer input, and just continue the current utterance. It can also slowly bring the system utterance to an end, by, e.g. changing the prosody, decrease the energy etc. so that the synthesized voice trails off at the end of a syllable, word or breath group. We call this behavior 'braking'. One precondition for this that the language generation module introduces braking indicators at word, phrase and sentence boundaries(cf. [4]). Another precondition is that the prosodic parameters in the speech synthesizer are accessible from the outside (cf. e.g. [5]).The most un-natural and annoying way is to just kill the synthesizer output.

Off-talk

In the vehicle, we have a free-speaking environment. Thus, another source of false alarms for barge-in, as well as a source of error for speech recognition is off-talk: people in the car speaking, but not to the system. Due to the space limitations and precedence of other installations in front of the driver's seat, standard microphone array technology can not focus narrowly enough, especially in the depth, to completely exclude off-talk, especially from behind the driver, to be picked up. We are therefore experimenting not only with better blind source separation, but also with different microphone arrangements in the vehicle. For this, we want to exploit the fact that the car, with all its problems, also is a 'known' environment in the sense that set positions, head positions etc. are relatively stable. A microphone for each seat in the car would eventually allow all the occupants to benefit from the ease an naturalness of speech dialogue, without interfering with other users. Furthermore, this would enable us to monitor the acoustic environment of the driver to possible detect distracting factors from the in-car communication of the occupants and to take appropriate measures to support driving safety.

Conclusion

In-vehicle dialogue requires sophisticated acoustic processing, both for the input and the output side. Some dialogue challenges, like intelligent barge-in, however, can not be solved by acoustic processing alone. We have shown that for these cases, the dialogue management needs to be closely coupled with the acoustic processing. The overall system behavior is what the driver, the customer, comes to experience. The task, therefore, is to optimize not only different components of the system, but to increase the co-operation and mutually enhance the performance by exploiting knowledge from other domains.

The vehicle, although known to be a 'hostile acoustic environment', offers a very good field for interdisciplinary approaches to enhance the performance of speech dialogue systems. For the researchers involved, over and above the solution of scientific and technological problems, it offers the additional motivation eventually contribute to increased traffic safety.

[1] Heisterkamp, Paul (2001): Linguatronic: Product-Level Speech System for Mercedes-Benz Cars. In: Proceedings of Human Language Technology (HLT2001), San Diego, Ca.

[2] Glass, Jim; Seneff, Stephanie (2001): Intelligent barge-in. In: Proceedings of Eurospeech '01, Aalborg, Denmark.

[3] Heisterkamp, Paul (1993): Ambiguity and uncertainty in spoken dialogue. In: Proceedings of Eurospeech '93, Berlin, Germany.

[4] Heisterkamp, Paul (1999): Time to get real: Current and future requirements for generation in speech and natural language from an industrial perspective. In: Becker, Tilman; Busemann, Stephan (Eds.): "May I speek freely?" Between templates and free choice in Natural Language Generation: What is the right NLG technology for my application? Proceedings of the NLG Workshop in conjunction with KI-99, Bonn, Germany, 1999.

[5] Poller, Peter; Heisterkamp, Paul (1997): A compact representation of prosodically relevant knowledge in a speech dialogue system. In: Proceedings of the Workshop on Concept-to-Speech Generation Systems, ACL'97/EACL'97 Joint Conference, Madrid, Spain.