Designing a Digital Companion That Increases Users' Trust in Vehicles Using Voice And Sounds

March 21, 2023

By Gabriel Haas, Senior Research Engineer

THE INCORPORATION OF HUMAN ELEMENTS INTO TECHNICAL SYSTEMS
Today's digital voice assistants have become immensely powerful and can help with a wide variety of tasks, starting from simple command execution to answering all kinds of broad questions. But even though they are incredibly helpful, they often can feel very technical, algorithmic, and rule-based – a reasonable perception given where the technology once was. This behavior of voice assistants can make them feel impersonal and not very empathetic. We feel that they lack the human touch as they are incapable of feeling and perceiving emotions.

As a UX researcher and interface designer, it is one of my goals to address and bridge this gap between machines and humans. We even have the right tools at hand: emotion recognition. Emotion-sensing technology is an important step in the evolution of artificial intelligence. It helps machines to understand and analyze human emotions to determine what kind of situation a user is in.

Emotion-based technologies can have a significant impact on increasing driver comfort, user confidence, and affinity for autonomous vehicles. Leveraging auditory emotion detection and emotion-aware dialog systems, we are working towards an empathetic and intuitive automotive assistant that can respond and give feedback to drivers with emotional intelligence and understanding, improving the driving experience.

A STRONG AND EFFECTIVE TEAM FOR THE BIG TASK
Automated vehicles are about to turn the automotive world on its head and bring unimagined opportunities to make better use of travel time. Simultaneously, there is still a low level of acceptance of this technology on a consumer level because they are afraid of losing control to immature technology.

A central prerequisite for the successful introduction of fully automated driving is the users' acceptance of the technology. Two decisive factors here are

- a sufficient degree of trust in the security of the system, and
- a noticeable added value or a positive user experience.

As part of the EMMI project, these two central components are addressed through the systematic conception, development and evaluation of an empathic human-machine interface.

While Cerence leads the development of voice and sound-based interfaces within this project, we are cooperating with a strong bunch of partners that contribute their strengths and expertise in visual driver monitoring, 3D avatars, display technology, and cognitive psychology.

As part of the EMMI project, Cerence created a Control Center that unifies a set of software components provided by Cerence. We refer to this control center as the Cerence Control Center (CCC). The CCC provides the following five key functions:

- Transcription of spoken language, Automatic Speech Recognition (ASR).
- Analysis and interpretation of the recognized words, Natural Language Understanding (NLU)
- Recognition and interpretation of paralinguistic information in speech and the emotional state of users (EMO)
- Creation and control of a dialog with the user (Dialog Manager)
- Synthetic generation of spoken language, Text-To-Speech (TTS)

EMMI

The illustration shows the interaction of the individual components and clarifies the individual steps that are necessary to provide an empathic, digital partner.

In order to connect the components and make information accessible and exchangeable between the project partners, a communication interface had to be defined and implemented. Via this interface, the partners Charamel, DFKI, and CanControls can retrieve relevant intermediate results of the system and process them independently. Every component in the EMMI project uses the same communication interface, thus enabling smooth communication between the components and partners.

AUTOMATIC SPEECH RECOGNITION & NATURAL LANGUAGE UNDERSTANDING (ASR & NLU)
The first key function is the recognition of the user's speech. This function is called speech recognition or ASR (Automatic Speech Recognition) for short. The speech recognition component in the CCC picks up the audio signal from the microphone, extracts the spoken language and converts it into text. For this purpose, language-specific models are used in the background.

The integration of speech recognition in the CCC was implemented in such a way that the microphone is always activated. A dialog can be initiated by the user with any wake-up word (WuW). Thus, the system continuously listens to hear if a phrase defined as a WuW is spoken and then interprets the subsequent words as input to the system. In addition, continuous listening is important for the emotion recognition component. Thus, it is possible to infer emotion from the incoming audio signal even if no voice command was spoken. But more about that later.

In the second step, the result of the speech recognition is passed to an NLU module. This analyzes the spoken words and a semantic representation is extracted from the text. For example, filler words can be filtered out and paraphrases for commands can be unified to a central user intent. For example, the instructions "Please drive me to Cologne" and "I would like to have a navigation to Cologne" can be unified to the same "intent." This

makes further processing easier since it is no longer necessary to pay attention to different wordings, but only the user's intent is used.

RECOGNITION AND INTERPRETATION OF PARALINGUISTIC INFORMATION AND THE EMOTIONAL STATE OF USERS (EMO)
The CCC's paralinguistic and emotion state analysis component offers the possibility of recognizing emotions and other paralinguistic characteristics (so-called traits) in the voice of the speaker through speech-based analysis. In the background, various voice characteristics are compared with specially trained, probability-based models in order to continuously make a decision about the current state. The values are not only available within the CCC, but can also be retrieved via the EMMI websocket interface, so that information can be passed on to other systems as required.

CREATION AND CONTROL OF A DIALOG WITH THE USER (DIALOG MANAGER)
A dialog manager is required to enable empathetic interaction with the user and to provide the user with the desired information. It receives the recognized "intents" of the NLU component and decides, based on these, how the dialog with the user should be continued. These options can be easily entered and extended with the Dialog Manager integrated in the CCC. It supports not only direct responses but also more complex sub-dialogs and can control any connected systems via a network interface. For example, the vehicle's window could be opened to let fresh air into the vehicle, or the lighting mood could be adjusted to create a calming effect on the driver.

The Dialog Manager thus forms the basis for intelligent conversation and is also the interface to all other connected systems, which both supply data to the Dialog Manager and perform the actions requested by the user.

THE SYNTHETIC GENERATION OF SPOKEN LANGUAGE, TEXT-TO-SPEECH (TTS)
In the Text-To-Speech component, the Cerence TTS engine can output words or sentences entered in the text field as computer-generated speech. It is also possible to adjust the volume, speed, pitch and timbre of the generated speech according to the user's preferences. This is important because the tonal character has a great influence on the perceived empathy of the digital dialog partner. It is also possible to have the language generated in German and English.

In cooperation with the consortium partner Charamel, it was decided that the creation of computer-generated audio files should be done as a service, which means that it can be easily queried and returns the results. Not only is the pure audio signal generated, but also Lipsync information so that a digital avatar can move its lips in sync with speech.

CERENCE CONTROL CENTER - A VOICE INTERFACE FOR EMPATHIC HUMAN-MACHINE INTERACTION
Voice interaction systems have been around for quite some time and have proven to be a very successful human-computer interface. They are a great way to interact with complex technical systems in a natural and human way, but there are still some challenges that need to be addressed.

One of the problems is the emotionality both in the speech output, within the dialog, and the recognition of the user emotion which has a great impact on the trust towards a system. Cerence is creating deep neural networks to recognize spoken language, detect emotions, and respond accordingly.

The CCC is a speech interaction system that enables us and our partners to interact with machines in an empathic way. It provides the building blocks, which in turn can be used by our partners to create and improve their own systems. All with the goal of being able to empathize with the user and thus build trust and make the interaction natural and intuitive.

To learn more about our work with EMMI, visit https://www.emmi-projekt.de/newsblog-en.html.