Customized voices are a great way to enhance and distinguish a brand

April 12, 2019

Personalization is one of the key features that distinguishes modern in-car smart assistants, mobility assistants or personal assistants from traditional infotainment systems as well as from one another. Building on white-labeled development platforms like Cerence Drive, car manufacturers have a chance to develop a personalized experience that is tailored to users' specific needs and enhances their experience to one that is uniquely theirs.

Sophisticated smart assistant solutions not only enable smart command and control functionalities but also dialogue-based collaboration. When users are having a real conversation with their systems, these systems not only need to understand the queries but need to be able to deliver an appropriate response –with respect to content and context but also voice, intonation, and style.

New advancements in deep-learning enable Cerence to offer a large variety of system voices that sound more human-like, can be customized according to customer demands and enable a great user experience across a large variety of day-to-day use cases. I sat down with Johan Wouters, Director of Product Management to talk about our latest advancements in deep-learning based text-to-speech, voice creation and customer benefits of these solutions.

Johan, what is the key rationale behind customizing the voice used in a mobility assistant?

Personalized custom voices are important features of a unique user experience. The first direct impression of a speech system is the quality and personality of the voice output. This is why we are offering customized voices to our customers. In addition to a unique set of capabilities included in the assistant, these voices are a great way to enhance and distinguish a brand. We are able to adjust voice characteristics like gender, age, timbre, accent, and speaking style. In addition, our natural language generation (NLG) can adjust the wording to reflect the message and driver context.

At CES we showed how our mobility assistant reacts to different user emotions detected by voice and facial expressions. Can you expand a bit on that principle?

We are convinced that this principle strengthens the emotional connection between the user and the mobility assistant. Users feel understood by the system and are more likely to establish a kind of emotional relationship to it. On the other hand, adjusting the text-to-speech output – like the messaging style and the sentence lengths, for example – can support or counteract different moods and improve the user experience. If the user is chatty, the system offers longer, more colloquial prompts. If the user is not, the information is delivered in a rather neutral and factual style. This system variation is based on real-time detection of the user's mood. But another possibility is adjusting the output based on learned preferences and context.

Let's start with user preferences: How can the user benefit from personalization?

Cerence's multi-style voices make it possible to render information in different ways. For example, if the system is providing a weather forecast, it can take into account individual weather preferences. Some people prefer cooler temperatures, others want to know about brightness, others are interested in wind speed, etc. There are many ways to derive personalized information, such as by learning it from interactions, asking specific questions, or having the user create profile settings. With our advanced embedded capabilities, the information can stay in the car and does not need to be uploaded to the cloud. Alternatively, the personalized data can be uploaded in a secure way so that it can also be used by a rental or shared car.

How can contextual awareness influence the voice output of the system?

When thinking about semi-autonomous driving and different transfer-of-control scenarios, the system could adapt the voice output depending on context. Again, we are talking about a scenario where the assistant is monitoring in real-time, but here, it is about fusing in-cabin and external sensor information. For example, if the system tracks that the driver's eyes are off the road and a transfer of control is required, the voice output could be more forceful and assertive.

Which additional features affect the user experience in everyday scenarios?

Cerence offers multi-lingual modalities which is important for a positive user experience when crossing borders or entering regions with foreign street names, etc. Multi-lingual capabilities ensure that landmarks, street names, etc. are pronounced according to the local phonetic rules and not according to the language set for the system. In addition, our latest research prototypes of human-like voices are almost indistinguishable from actual human speech. I brought some examples and bet you won't be able to get all of them right.