• Lack of user expectations: Few people have used many multi-modal speech applications. Just as in the early days of speech-enabled IVRs, there is a general lack of expectation about how multi-modal speech applications should work. Priming users with appropriate language constructions (for example, with sample text in input fields) is crucial.

• Centralized recognition with high-quality audio: In mobile speech applications, the audio signal is typically sent over the data channel. This enables high-quality (rather than telephone-quality) speech to be used, improving accuracy. Further, recognition is often done "in the network" (rather than on the device) – enabling utterance and usage data to be logged and used to improve service quality.

Adding speech input to mobile applications is becoming more common: examples include Google App, Vlingo, Cha-Cha, Siri, Speak4it, and others. Moreover, a feature in some versions of the Android OS add a microphone to the on-screen keyboard, allowing a user to speak instead of type at any point. There is a clear opportunity for speech on mobile devices. Ultimately, speech will be a successful addition to mobile applications if it improves the user experience. There is a clear opportunity for AVIxD to apply and adapt its knowledge of speech, human behavior, and design to this growing space.

David Suendermann

david@speechcycle.com

Five techniques multi-modal apps (should) inherit from speech science

Throughout its century-old history, speech and language science generated a huge scientific heritage with a sheer endless number of mathematical models, algorithms, and techniques. Thousands of scientists have spent decades of their valuable lifetime to make machines recognize, understand, react to, produce, translate human speech, and now you guys claim that ``newborn babies are handed smart phones'' rather than taught how to speak? As a speech scientist, I should be worried about the prognosis that ``speech is dead''.

So, what can we do to make us indispensable in the age of multi-mode? Can we possibly use the accumulated speech knowledge and reshape, transform, multi-modalize it? Let me give five examples speaking by themselves to this question.

- UNIT SELECTION is a technique invented in the 1990s to produce high-quality speech synthesis by selecting and concatenating segments from a prerecorded speech corpus. While unit selection became the most popular speech synthesis technique, recently, it also started a career in the visual mode where it is applied to produce photo-realistic talking heads (aka avatars).

- CONTENDER (academically aka reinforcement learning) is used in spoken dialog systems to train optimal decisions and parameters by randomly routing production traffic to different scenarios and settings and evaluating the success of these routes using a global reward function. Decisions and parameters are iteratively adjusted to maximize the system's overall reward. This approach is particularly useful when an interaction designer has doubts about what the optimal implementation of a certain task is. For instance, the designer may not be sure of what prompt to use, which recognition settings are best, whether to use directed dialog or open speech, etc. The more free variables there are, the harder it gets to make manual decisions.

In the multi-modal world, the number of free variables increases significantly. Now, we need to decide upon what to present as speech vs. on-screen, where to enable speech recognition, how to implement confirmation, n-best recognition results, etc. The space of possibilities is so immense that there is a strong need for computer assistance to multi-modal interaction design, and contender has proven to be a powerful technique for optimal decision making in high-trafficked applications.

- N-BEST LISTS are well known from speech recognition, machine translation, text search, and so on. They are lists of recognition, translation, or search hypotheses ordered by descending confidence. In spoken dialog systems, sometimes, n-best results are used in the scope of belief systems that keep a multi-dimensional space of system and user states and only make hard decisions if ultimately required. The avoidance of hard decisions has proven to result in higher resolution rates than conventional approaches.

A light-weight implementation of n-best lists uses less likely hypotheses as a backup response in case of a rejected confirmation to save another collection. An example dialog reads

A: How may I help you?
C: I have no picture.
A: It sounds like you're missing some channels, right?
C: No.
A: My mistake. Is your picture completely out?
C: This is correct.

N-best lists can be similarly applied when it comes to multi-modal applications. For instance, after a speech input, instead of confirming in the above fashion, one can use a visual display to show the n best recognition hypotheses. Now, it is up to the user to select the best match from this list very much like one does during a web search. This approach does not only shorten the conversation by avoiding multiple confirmation steps, but it also increases the overall recognition accuracy. While the recognition accuracy of the first best hypothesis of a large-vocabulary open speech input rarely exceeds 90%, it quickly approaches 100% when more and more hypotheses are added to the list.

- ENGAGER is an algorithm that exploits the fact that the order in which questions are asked or actions are performed in spoken dialog systems can have a significant impact on the average number of questions asked or actions performed in the course of a dialog. As Engager is a generic algorithm dealing with information entities, it is not limited to speech but can be applied to each input modality. Questions can be answered using speech, touch input, back-end services, gestures, or whatever one can think of.

- LANGUAGE MODELS are basic component of many speech and language processing disciplines. They provide probability estimates of word sequences that are required in large-vocabulary speech recognition, machine translation, spoken language understanding, and so on. Apart from the speech part of a multi-modal application, also other input modalities such as text input using standard telephone keys or the touch screen of a smart phone make use of language models or similar statistical models. For instance, predictive text technologies such as T9 for cell phones or virtual keyboards for smart phones are based on this kind of statistical models.

These five examples show how speech and language technology is expanding its scope towards other input and output modalities and, in doing so, provides a valuable tool set in the age of multi-mode.

Charles Galles
charles@cgalles.org