Int J Speech Technol (2007) 10: 163–166
DOI 10.1007/s10772-009-9027-7

Editorial

Amy Neustein

Received: 30 March 2009 / Accepted: 1 April 2009 / Published online: 16 April 2009

1 "*Bringing the Mountain to Mohammed*": the speech industry's effort to accommodate the user by improving design and assessment of voice-user interfaces

In a past issue of the *Journal* (Vol. 9, Nos. 3–4), fourteen contributors, drawn from a wide array of specialty areas in speech technology, presented viable approaches to some of the greatest challenges facing spoken systems today:

(1) Designing speech synthesis systems to perform optimally in high stress interactions;
(2) Restructuring hierarchical voice menus to improve efficiency and usability of telephony interfaces;
(3) Speeding up the process of ASR (automatic speech recognition) applications without compromising accuracy rates;
(4) Perfecting natural language understanding for speech interfaces used in critical systems;
(5) Skillful balancing of training and testing conditions for (medical) dictation systems to learn how to properly evaluate speaker adaptation to such systems;
(6) Bridging the cultural chasms in ASR technology so that non-English languages, such as Arabic, may benefit from advances in user-centered speech interfaces; and
(7) Controlling for degradation in the performance of Tamil speech recognition systems by using both TSM (time scale modification) and VTLN (vocal tract length normalization) techniques.

A. Neustein (✉)
Linguistic Technology Systems, 800 Palisade Avenue,
Suite: 1809, Fort Lee, NJ 07024, USA
e-mail: amy.neustein@verizon.net

While each of these contributions to that past issue of the *Journal* focuses on a different facet of speech application and design, the contributions are noticeably alike inasmuch as they all share a "user-centric" approach to building speech interfaces. The contributors are not alone in this *Weltanschauung*. In fact, much of the speech industry—despite the rapid advances in speech architectural design and the likely unwieldiness found at early phases of implementation—remains "user-centric." This editorial—prepared specially for the *Journal*—is intended to draw attention to three major areas of speech application design: automated directory assistance; telephone-based spoken dialog systems; and voice search in mobile applications. Particular attention is paid to how companies at the forefront of these technologies have been at pains to make everyday deployment of voice interfaces user-friendly.

2 AT&T Labs designs benchmark tests of directory assistance services

In the last issue of the *Journal* (Vol. 10, Nos. 2–3) AT&T Labs' researcher Harry Chang presented a "series of benchmark tests" to evaluate how machine models for automating directory assistance *compare* with human (operator) performance in answering callers' directory assistance requests. As Chang's article stated, in the last decade "one of the most active fields in ASR-driven applications in the U.S. telecomm industry is the automation of operator-based DA [directory assistance] services." Yet, in spite of the pervasive use of automated DA services, the systems deployed today, according to Chang, "still perform at a level that is far below most of the trained human operators employed by major telecomm service providers in the United States."

For example, in a series of benchmark tests—configured specifically for DA-related tasks to evaluate the performance of state-of-the-art and commercially available HMM (Hidden Markov Model) based ASR technologies—it was demonstrated that "the best [automated] system achieves a 57.9% task completion rate on the *city-state-recognition* benchmark test" and a "40% task completion rate for the *frequently-requested-listing* benchmark test."

Such sobering statistics on the real performance of automated DA systems present a major challenge to the speech industry. That is, to narrow the performance gap between ASR-driven machine models and human operators, an ASR-driven DAA (directory assistance automation) platform must complete the task of answering the caller's DA request in less than 20 seconds—the accepted industry standard for the average length of time for human operators to answer callers' DA requests. Nevertheless, as Chang pointed out in his *Journal* article, techniques for "disambiguation and error recoveries that have been proven to be effective for ASR-driven Interactive Voice Response (IVR) systems"—used in call centers today—are "not easily transferable to the dialog designs" of automated DA services, because those techniques, requiring many (speaking) turns of dialog to perform adequate error recovery, are more than likely to exceed the standard 20 second time limit used by human operators to fulfill a DA request.

Notwithstanding this daunting reality, Chang does not back away from the challenge of designing and implementing ASR-driven DAA platforms to emulate human operator performance. By "comparing machine performance with that of the trained human DA operators on a dialog-by-dialog basis," the AT&T Labs' research team is able to "precisely measure the performance gap at *each* step, and gain a better understanding of *why* and *where* the current ASR-driven DAA platforms fall short relative to the same tasks fulfilled by their human counterparts."

To better understand the performance of human operators, AT&T Labs' research protocol entails close study of how directory assistance operators "handle hundreds of different dialogue paths" with great efficiency. Given that training of human operators continues to advance (e.g., DA operators can now search automatically-expanded geographic areas for unclear phone listing requests), Chang predicts that "more studies are required ... to truly understand the techniques used by experienced DA operators to achieve the level of [human] performance, which may take decades for a computer-based cost-effective DAA platform to match." Nevertheless, AT&T Labs' research into the performance of an ASR-driven DAA platform is further evidence that by identifying the *specific* factors associated with the performance gap between human operators and machine models, we move closer to designing speech interfaces that truly emulate human-level performance.

3 SpeechCycle designs and tests an innovative metric for evaluating spoken dialog systems

Roberto Pieraccini, CTO at New York-based SpeechCycle, and member of the *Journal's* editorial board, has taken on the Herculean task of designing and testing a useful new metric that serves as a single, subjective numerical rating to evaluate the performance of spoken dialogue systems—across different populations of users and subject domains.

In a paper presented at the *2008 IEEE Workshop on Spoken Language Technology*, Pieraccini and his team of researchers described their new metric for evaluating Caller Experience (CE) as distinguished from other IVR performance metrics for two reasons: First, "it is a subjective, qualitative rating" of a user's experience with an IVR system. Second, the rating is "provided by expert, external listeners and not the callers, themselves." Pieraccini and his co-authors astutely point out that using "external listeners"—a process that may be automated likewise to reduce human labor costs of hiring expert listeners to rate calls—can prevent much of the skewed research findings that occur with a self selected sample of users and/or where there is no uniformity in the way each subject interprets the survey question.

As call centers increasingly adopt spoken dialogue systems to handle customer service requests, better metrics for evaluating CE have become *sine qua non* to the functioning of such IVR-driven call centers. SpeechCycle's robust and reliable new metric—using data from 1,500 calls annotated by fifteen expert listeners—is useful for monitoring dialog systems in deployment and for identifying individual problematic calls in which the system under-performs, as in cases where the system fails to accurately identify and satisfy the reason for the call, or where the system fails to recognize altogether what the user said. Such feedback is critical to any well-functioning IVR-driven customer care and contact center, as noted by Pieraccini and his research team: "Knowing how callers seem to experience a system can help guide business and design decisions. Specifically, a call's CE rating can indicate which interactions need to be streamlined, simplified, or made more robust."

SpeechCycle's innovative CE metric is an example of sharper methods now in development for evaluating IVR performance—methods that, with scrupulous design and testing, promise to provide more effective and timely handling of caller requests. This is good news for any user who at one time or another has found himself/herself trapped in the "maze" of an IVR system.

4 Vlingo develops a flexible voice-user interface to adapt to users' individual speech patterns in mobile search applications

In a recently published white paper on *unconstrained* speech recognition, Mike Phillips, co-founder and CTO of

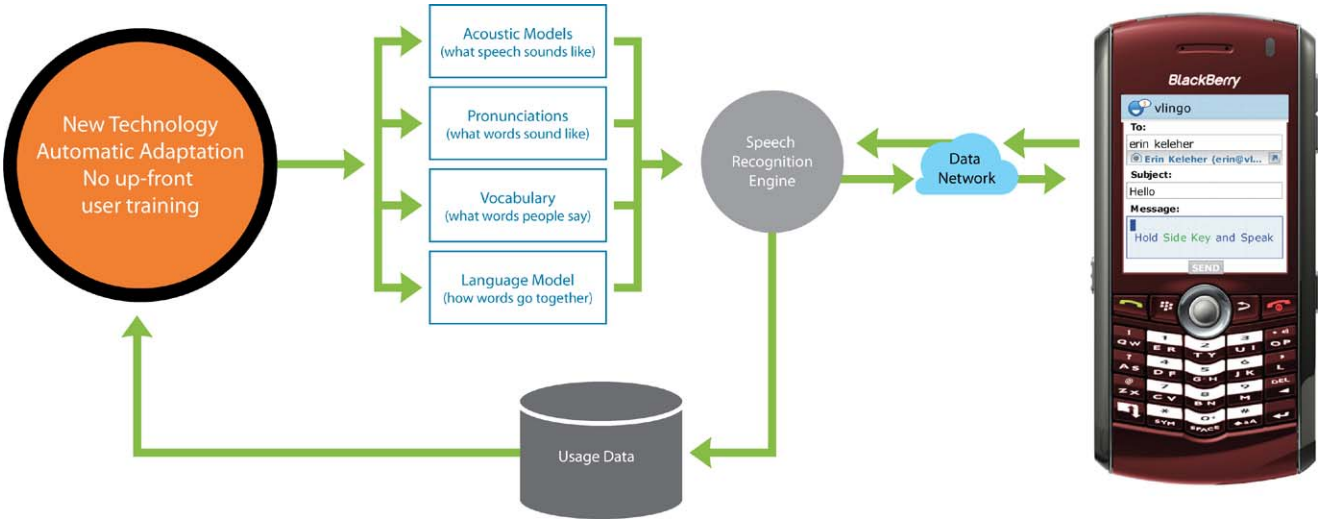

Fig. 1 Vlingo Adaptation Architecture. The core Speech Recognition engine is driven by a number of different models, each of which is adapted to improve its performance based on usage data

Vlingo (Cambridge, MA), described his company's development of a novel voice-user interface—using a very large vocabulary—to solve "the overall usability challenge on mobile phones." Phillips explains how *adaptive* Hierarchical Language Models (HLMs), "which are based on well-defined statistical models to predict what users are likely to say given the words they have spoken so far," are better able than constrained grammars to perform voice-driven web search in mobile applications. Speech recognition systems, using the latter approach, explains Phillips, are severely hampered when users employ "out of grammar" words and word phrases, which is understandably "the most frequent cause of all misrecognitions." Phillips opines persuasively that this is why "traditional speech implementations [using constrained grammars] have not always been warmly embraced by users."

In an interview last May with *Speech Strategy News*, a pre-eminent industry newsletter published and edited by speech industry savant Bill Meisel, Phillips explained that "by supplementing the core speech technology with this new approach," HLMs can "adapt to the user's habits and to the specific text box in a particular application, increasing [speech recognition] accuracy over time."

In his company's white paper, Phillips expounds on what is meant by "adaptation":

- **Adaptation:** In order to achieve high accuracy, Vlingo makes use of significant amounts of automatic adaptation. In addition to adapting the HLMs, the system adapts to many user and application attributes: for example, learning the speech patterns of individuals and groups of users; learning new words; learning which words are more likely to be spoken into a particular application or by a particular user; learning pronunciations of words based on usage; and learning peoples' accents. The adaptation process can be seen in Fig. 1.

At this point the reader may be wondering how an HLM, built on a very large vocabulary—a lexicon that naturally grows as the "open voice" interface learns the user's speech patterns—can be scaled for a mobile environment. The answer, according to Phillips, lies in sever-side processing: the Vlingo deployment architecture uses about 50–90 KB on the mobile device, while communicating over the "mobile data network to a set of servers which run the bulk of the Vlingo processing." This networking design approach—Vlingo's servers are designed to respond to users' requests in less than half a second—"enables the use of large amounts of CPU and memory resources needed for unconstrained speech recognition." Most important, Phillips explains, this network design allows a flexible adaptation of a voice-user interface—that is, adaptable to any platform or domain—in mobile applications across a wide population of users. Vlingo's voice-enabled applications include a GPS-enabled navigation for local business search and a mapping application; a Voice2Txt—a text messaging application with threaded conversation similar to instant messaging; a web search application that allows a user to search the web by voice; a video viewer that searches and plays YouTube videos; and other voice-powered applications.

Against this sanguine picture of a highly adaptable, high-functioning voice user interface deployed in a variety of voice-enabled mobile applications—across wide populations of users and diverse subject domains—one must weigh certain caveats. For example, a user might refrain from speaking his/her search request if other people are nearby and the user's need for privacy is paramount; or where high noise and other acoustic factors compromise the speech

recognition; or where there is a lack of network coverage in the area. To combat such contingencies, Vlingo has built *multi-modal capabilities* into its user interface so that, as explained by Phillips, "the user can freely mix keypad entry and speech entry—at any time the user can either type on the keypad or push the 'talk' button to speak." This correction interface can "allow the user to correct the words coming back from the speech recognizer." Among the many options, "users can navigate through alternate choices from the speech recognizer (using the navigation buttons), can delete words or characters, can type or speak over any selected word, and can type or speak to insert or append new text wherever the cursor is positioned."

Vlingo is right on target with its design of a flexible user interface, which permits adaptation of the mobile search device to the constraints of the user at any given moment. At the same time, there are clearly so many areas in which improvement is needed that it's hard to say which of the exciting developments sketched here will prove most important or most popular. Perhaps the best prediction about where speech-enabled mobile devices are headed comes from Bill Meisel, president of TMA Associates and a member of the *Journal's* editorial board. In an industry press release on this year's Voice Search Conference (a conference co-sponsored by AVIOS and TMA), Meisel summed up user behavior vis-à-vis the speech industry: "Once one adopts any voice-enabled application, one has essentially learned the voice user interface. It's easy to predict that adoption of voice input on mobile devices will grow quickly, but it's more difficult to see how any one company can come to dominate."