

VUI Visions

Voice User Interface Design: From Art and Science to Art with Science

Roberto Pieraccini and Phillip Hunter, SpeechCycle

*In this guest column, we ask designers skilled in creating Voice User Interfaces to highlight a particular aspect of VUI design inspired by actual deployments. In this issue, Roberto Pieraccini, Chief Technology Officer, and Phillip Hunter, Vice President of Interaction Design, **SpeechCycle**, discuss how alternatives in parts of a VUI dialog can be chosen objectively based on statistics compiled during use of the application. Pieraccini has worked at CSELT, Bell Labs, AT&T Labs, and IBM T.J. Watson Research, led an R&D team at SpeechWorks, and written more than 100 papers and articles. Hunter has designed scores of applications, built successful teams, led Fortune 500 projects, won awards, and invented processes and tools at several previous companies.*

ART HISTORY

Through the years there has been a tendency to consider the design of voice interactions and voice user interfaces (we will refer to both as VUI) primarily an art. The ability to create aesthetic and functional experiences was made possible and mediated mostly by the knowledge gained in having built dozens of applications. The evaluation of VUI solutions was based by and large on the introspective ability of the designers and their capacity to predict consequences and to make the right choices. Challenging and new situations required consulting with other designers and, more often than not, several possible solutions were discovered, not counting those proposed by the customers and by anyone else on the project with a self-determined knack for VUI.

However, this view of the art of VUI only partly parallels the applied education and craft found among practitioners of the classic arts. In those fields, science is also a tool and even a source of inspiration. For example, the learned sculptor understands how to wield force in cooperation with the characteristics and limits of steel and stone. But, along with other arts, there is a singular, common, and hard-to-know component in the final product: the emotional response produced in those that experience it. Yet, that causes the artists to know and master their tools and domain even more so they can understand, gauge, and even manipulate the response their art provokes. So, art and science are companions in the pursuit of excellence and desired effect, complementing and even relying on each other. However, while VUI producers have long used the tools and techniques of the trade, rarely has there been a full use of science and an application of what can be learned from the responses provoked by the voice system. This ignorance of the data and its meaning has been a shortcoming and even a failure point for many speech deployments.

THE AGE OF SCIENCE

From its very early stages, SpeechCycle has adopted a scientific approach: choose the design—or designs—that optimize well defined measurable criteria such as, for instance, automation rate and average handle time. And that is done in an objective manner by using statistical evidence on large amounts of live data. This is especially important as speech applications grow in scope and complexity. Calls to SpeechCycle applications involve dozens of turns and can follow hundreds of dialogue paths. [Editor's Note: SpeechCycle announced in April that it had handled its 50 millionth technical support call for cable customers (SSN, May 2008, p. 25).]

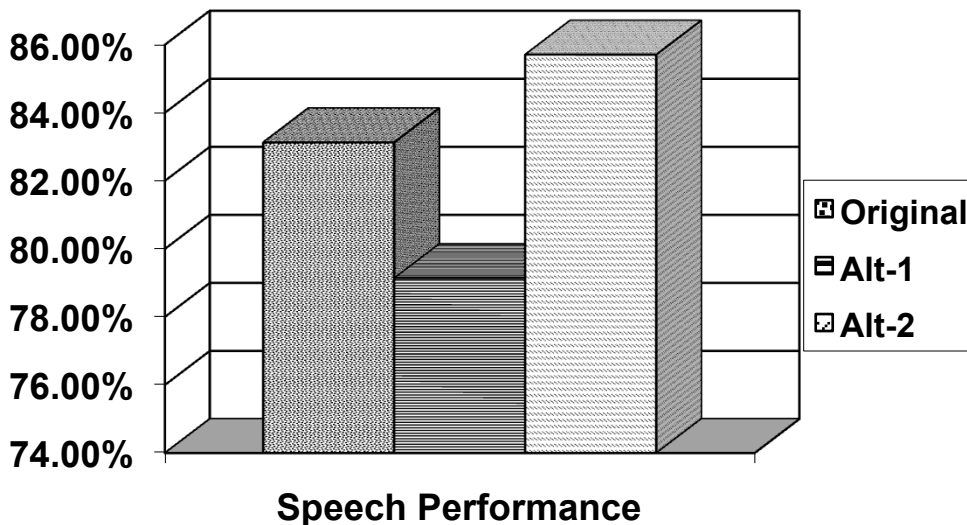
The idea of making design choices based on the statistical optimization of objective measurable criteria is not new. The approach of using machine learning—reinforcement learning in particular—for the optimization of dialog systems was first introduced by Esther Levin and Roberto Pieraccini in 1997 [1] while at AT&T Labs, and then extended and improved by other researchers such as,

among others, Steve Young [2] at the University of Cambridge, UK, Olivier Pietquin [3] at the University of Mons, Belgium, and Oliver Lemon [4] at the University of Edinburgh. For a current discussion on the implications of reinforcement learning for commercial dialog systems see a recent article by Tim Paek (Microsoft) and Roberto Pieraccini (SpeechCycle) published in *Speech Communications* [5].

Essentially reinforcement learning uses algorithms that favor the choice of design alternatives that increase the average automated agent reward defined as a function of the objective criterion to optimize (e.g., average automation rate, call duration, number of no-matches, or a combination of them). While the application of reinforcement learning to the full design of dialog systems, although possible [1], is still impractical [4], SpeechCycle adopted a partial design approach where only a small but significant number of *exploration* points are set in a call-flow. Each exploration point corresponds to two or more reasonable design alternatives. Exploration is implemented by a feature—called “Contender”—of the SpeechCycle RPA (Rich Phone Application) Express platform.

CASE STUDY I

As a simple example to illustrate the power of statistical exploration and analysis we show the use of this model for the optimization of the portion where the reason for the call is captured within a large technical support application deployed for cable TV providers. This part of the application was underperforming and we sought an increase in the numbers of callers successfully giving a call reason. Two new structures, “Alt-1” and “Alt-2”, were determined to be viable candidates and both were placed into the application in parallel to the existing “Original.” The Contender feature of the RPA platform randomly selected one of the three options for each incoming call using a uniform probability distribution. The selected option was marked in the logs of each call. After about 26,000 calls, the differences in the automated performance between the subsets of calls corresponding to each selected option were deemed to be statistically significant based on a standard significance test. Additionally, we examined a number of typical measures, like grammar accuracy, hang-up and opt-out rates, number of successful problem captures, and number of overall successful calls. The results clearly indicated that Alt-2 was the better approach with relative reductions in opt-outs (18%), hang-ups (23%), and improvements in speech recognition performance.



So, while VUI art was certainly needed to create the dialogue structures and prompts used for this experiment, in the end it is also the data-driven scientific method that determined which approach was best and should be used in full production.

CASE STUDY II

Even when competing VUI designs are not being tested, SpeechCycle constantly and thoroughly analyzes large amounts of data produced by caller interactions. Data mining and analysis, mostly performed with the help of automated tools, results in useful information about caller responses that influence the evolution of the application. For instance, the analysis of out of grammar (OOG) responses is particularly interesting for the improvement of the system performance. Often callers give adjunct information or answer a question that has not yet been asked. As humans, we do this frequently:

Person 1: "Where's your car parked?"

Person 2: "I took the bus today."

Traditionally, spoken dialog systems crash and burn with those kinds of interchanges, despite them being a stalwart conversational aid. However, with the help of data analysis and good design practices one can revise the system and properly handle such utterances, as in the following example:

IVR: Which one can I help with? Just say "my bill", "tech support", "orders", or "appointments".

Caller: "Internet service"

IVR: Okay, internet service. Just say "my bill", "tech support", "orders", or "appointments".

Caller: "tech support"

For the above example, significant amounts of data indicated that such utterances as "internet service" were common even though the directed dialog prompt did not mention them as a possible choice. Design analysis indicated, of course, that the utterances were in context for the application. Thus, SpeechCycle designers created the simple method of handling them that is illustrated above, which allowed callers to flow through the interaction smoothly and successfully. The detection, analysis, design, and implementation of this instance took about 40 person-hours over three weeks. Without the combined approach of data collection, analysis, tools, and expert VUI design happening in a very timely fashion, such a solution would likely have occurred several months after deployment and would have depended, in traditional tuning, on whether a speech scientist happened to notice the pattern. Traditional tuning performed sporadically offers only haphazard responses to often stale data. On the other hand, constant evaluation and analysis can lead to remarkable short-term improvements.

CONCLUSION

A better partnership of data and design, using careful measurements and automated continuous analysis, can ensure that appropriate application evolution occurs promptly. It is this reliance on scientifically approaching performance evaluation and design decisions that drives the next evolutionary step for all speech applications. Art? Yes, but not an art without scientific discipline. SpeechCycle has shown that scientific assessment and analysis can effectively guide and evaluate the art of VUI design.

REFERENCES

1. E. Levin and R. Pieraccini, *A stochastic model of computer-human interaction for learning dialogue strategies*, Proc. of Eurospeech 1997, Rhodes, Greece, September 1997.
2. S. Young, *Talking to Machines (Statistically Speaking)*, Proc. of 2002 International Conference on Spoken Language Processing (ICSLP), Denver, Colorado, September 2002.
3. O. Pietquin, *A Framework for Unsupervised Learning of Dialogue Strategies*. Presses Universitaires de Louvain, SIMILAR Collection, 2004.
4. O. Lemon and O. Pietquin, *Machine Learning for Spoken Dialogue Systems*, Proc. of Interspeech 2007, Antwerp, Belgium, August 2007.
5. T. Paek and R. Pieraccini, *Automating spoken dialogue management design using machine learning: An industry perspective*, Speech Communication 50 (2008), pp. 716–729.