ENTERPRISE INTEGRATION GROUP

SPEAKER IDENTIFICATION & VERIFICATION

A Position Paper by:

Bruce Balentine,
Enterprise Integration Group
bruce@eiglabs.com

Introduction and Overview

This document is a position paper prepared by Bruce Balentine of Enterprise Integration Group (EIG) for the Workshop on Speaker Biometrics and VoiceXML 3.0 hosted by SRI International, Menlo Park, California, 5-6 March, 2009. The paper touches on concerns expressed by EIG clients about:

- User acceptance of enrollment;
- Challenge versus security questions;
- The true cost of security and privacy violations in the phone channel;
- User preference for long versus short enrollment;
- More or fewer questions and the effect they have on user trust; and,
- Enterprise preference for certain challenge questions.

I and EIG Labs have been studying these concerns—looking for a well-integrated approach that makes the overall user challenge more tractable. We hypothesize that we want to:

- Find a way to avoid the problems and doubts altogether;
- Generalize and simplify the core behavioral issues;
- Find a way to fix problems without having to understand them; and,
- Use abstraction to bring disparate elements closer together.

A basic thesis revolves around flexibility in the supporting integration tools for the various media that may be combined to produce the end results desired. I would like the opportunity to discuss these thoughts at the SIV workshop, and to explore ramifications that might impact the ideal API designs for various media, the voice biometric features supported by VoiceXML 3.0, and the support for plug-ins that might exploit horizontal as well as vertical integration of technologies provided by different vendors.

This paper does not propose solutions—it is intended only to describe the problems from a user interface designer's perspective.

The Challenge of Uncertain Media

Some media are deterministic or "certain" in their behaviors. A bistable switch, for example, can be thought of as existing in one of only two possible states. Although there are "in-between" positions that must be considered by button designers,[1] the indeterminacy of these positions is not carried upward to the high-level user interface designers that make decisions relevant to conscious human performance. That is, user interface designers need not consider superposed states, false acceptance, and/or false rejection when designing applications that make use of bistable switches.

Other media, in contrast, are inherently non-deterministic or "uncertain" in their behaviors. Media that make use of statistical math tend to fall into this category. Uncertain media exhibit many possible states, some of which are the "in-between" states so undesirable for bistable switches. Automatic speech recognition (ASR) and speaker identification and verification (SIV) are both non-deterministic media. Even high-level user interface designers must consider class-overlap rejections—as well as true-and-false acceptance and true-and-false rejection—for every state of the interface.

Uncertainty complicates design philosophy by requiring attention to a multiplicity of conditions that cannot be classified with certainty. In particular, deterministic concepts such as "correct or incorrect," accuracy, error, error-recovery, and "robustness" are often misleading when one designs for uncertain media. Inappropriate assumptions and false mental models become confounding with such media.

[1] Button designers must employ de-bouncing circuitry, mechanical or electromagnetic hysteresis, and similar persistence techniques to make a switch bistable. But these techniques are invisible to the human user operating at larger scales of space and time. To the user, the switch is on or off.

Lessons Learned from ASR

From its earliest years, the inherent uncertainty of speech was ignored by speech vendors focused on simplifying their message to buyers. The design philosophy (now known to be flawed) that emphasized "accuracy" over all other measurement parameters was especially pernicious, because it led to fundamental flaws in API-design, error-recovery-dialogue, and logging-and-reporting decisions. These flaws plague the ASR industry to this day.

The core misconception that led to these decisions can be characterized as a *refusal to embrace uncertainty*. "I don't want to deal with uncertain conditions in my user interface input," the designer might say. "Therefore, I will treat the non-deterministic medium as a *not-yet-mature* deterministic medium. Errors are not my fault, they result from the current immaturity of the classification technology. As the accuracy of the ASR rises, I will experience less and less embarrassment when my interface gets lost in uncertainty."

Designs based on this apparently reasonable paradigm led to error-amplification effects when perfectly natural user behaviors triggered speech recognition "errors" that were then mishandled by clumsy and rigid "error-recovery" dialogues.

My position in this paper is that the industry promoting SIV as a technology appears to be committed to this same flawed track, and is likely to suffer similar customer-acceptance problems unless key industry stakeholders adopt a more flexible model.

Signs of the Flaw in Thinking

When visiting customer sites, I hear stakeholders argue over an equal error rate (EER) of 2% versus 0.5% as though the numbers are fixed and intrinsic to the current state of technology. I hear security experts declare that, "point-five percent is nowhere near good enough, we need five nines before we'll consider this technology." I hear about bake-offs and plans for future accuracy improvements. I see pilots that are derailed when errors—an inevitable consequence of uncertain media—occur. I see projects started and stopped, and I see sluggish uptake from the buying community despite claims that "the security and privacy problems plaguing my call center are acute."

The above are symptoms of cognitive dissonance, and they represent a public presentation of the flaws in thinking that dominate uncertain media.

Letting Data Interact

One important way to deal with uncertain media—indeed, one of the ways that human beings make decisions based on uncertain data—is by comparing data from a multiplicity of sources and then letting those data interact. In speech recognition, for example, application performance can exceed "raw accuracy" when multiple states are considered (including the history of previous sessions with this user), when turn-taking confidence is combined with ASR confidence, and when the timing of input is tracked across the dialogue. It is well-known that relying on ASR to make a guess about *what the user just said* can operate in concert with an SIV technology guess about *who the user probably is* to increase the certainty that the ID&V process is under control and reliable.

By extension, we at EIG feel confident that similar uncertain guesses—whether the user is lying, whether the input is background noise or user speech, whether the user is resorting to reference materials rather than speaking from memory, whether any derogatory flags should be dropped into a customer record—can be teased from input and context data to produce increasingly "certain" decision about machine behavior. The key is to find ways to let these multi-leveled data interact in some sensible and properly-weighted way.

There are many approaches today to letting data interact, and new ways will appear in the future. Most of them entail statistical comparisons of disparate probabilities, for example:

- Fuzzy logic
- Neural networks
- Genetic algorithms and selectionist models
- Brute-force algorithmic approaches

A "Great Simplification"

ASR, SIV, and the Telephone Channel

All information in the phone channel is exchanged in a bi-directional series of interactions. The machine presents information via voice, and the user responds with touch-tone or speech. Snippets of information are separated into turns with intervening clarification and error-recovery exchanges. Designers—and sometimes even users—are expected to keep up with the arbitrary classification of these snippets. In a speech API, for example, the terminology associated with events is biased:

- Some returns are “correct” responses.
- Other returns (timeout, OOG, rejection) are “errors.”

When a high-confidence recognition is returned from the ASR, this result is presumed to be “correct” despite the (usually rare) occurrence of substitution errors and the (more frequent) occurrence of OOG false acceptance. When timeouts and OOG are reported, the assumption is that these are “errors” in the sense that errors occur in a deterministic medium—leading to dialogues that are misleading or downright wrong.

SIV technology, similarly, is easily interpreted by designers through arbitrary classifications:

- Some dialogues are for "enrollment," while others are for "verification."
- Some questions are "challenge" questions, while others are for "security";
- We need more enrollment samples in some cases than others.
- "ID" is somehow different from "V."

The Three Players

In all call center interactions, we have three players—a caller, an IVR, and an agent. The totality of contact center interactions is handled through some mix of these three players. Each of these players has responsibility for enforcing enterprise policies. These policies are essential for fraud-detection and -prevention, but are also important for personalization and customization (and therefore effectiveness and efficiency) of the user experience. They are also important in guarding the legitimate privacy concerns of all partners in the dialogue.

The precise mix of these problems, their cost to end user and enterprise, and the degree of minimum intervention required all vary from instance to instance, and also over time for any given enterprise. What is needed is one central place for the specification of the enterprise policy as it changes over time, and one central repository for all of the partial snippets of information that together, over time, provide the application with adequate knowledge about what’s going on in the user’s world. Snippets of information include .wav files with known or unknown content, voiceprints, calling patterns and history-lines, mined data, human-specified flags, and other factoids that appear during the course of repeated interactions between an individual human and the enterprise with which he or she interacts.

In the end, a telephone-based application simply asks questions. The questions may be about the claim of identity—asking the user basically to "sign in" before going to an agent. In other cases, the questions allow the application to invoke its power to deny access or to re-route the call. Actual machine behaviors are based on corporate policy which is constantly changing. Sometimes, an SIV sentry is merely "sniffing around" and not explicitly present, for example

during a conversation between caller and agent. Sometimes the application may intervene more directly—as a part of the IVR self-service dialogues, or sometimes when an agent chooses to "transfer" the call to the IVR for some specific reason. It is part of the ergonomic design of the application that the reason for asking questions, the rights and responsibilities of the caller, and the expected (required) reaction of the caller to the sentry be made clear.

However, the caller need not always know everything about the sequence of events or the underlying technology. Sometimes the IVR may ask a question (or two or three), only to record the answer and save it for future use. This recording may be whispered to the agent for subsequent analysis or validation, or may be tagged by downstream events as uncertain answers become more certain over time.

Agent interaction of future events may identify the contents of these recorded files, allowing them to be tagged and applied to various uses (including biometric enrollment). In other cases, the recording proves later to have no value and is deleted. Sometimes the sentry will require explicit recitation from the user, other times it's a multiple-choice or yes-no question. Sometimes it's required, sometimes it's voluntary.

The Goal

All of these are options to the application designer if the software tools support delayed decision-making and cross-technology data interactions. The ideal application stores everything safely. It learns as it goes. It is an entity that behaves according to enterprise policies that are easily decided and changed. It can be incorporated into a freestanding box on either side of the IVR, or it can be incorporated into the IVR itself as a leg of the dialogue. It may also monitor the agent-caller interaction if the call center architecture supports that model. It doesn't always get the whole answer and a human agent must finish ID&V, but over time more and more calls are tightened and secured. Eventually, there is a high probability that a given caller will be confidently ID&V'd repeatedly and with minimal hassle, but that will only happen over time, after a series of low-risk individual steps.

Result: interactions are more secure, personalized, private, and focused.

Conclusion

In this model, we don't need to know how much we're preventing fraud or how much fraud we have now. We don't need to know whether personalization or privacy is a greater incentive to this or that caller. We don't need to know what the instantaneous accuracy of a given biometric event is. Instead we measure throughput, and balance the load between IVR and agent (the easiest ROI to understand) and everything else emerges spontaneously from the behaviors of the system as a whole. I would like an opportunity to discuss this design

iors of the system as a whole. I would like an opportunity to discuss this design philosophy with the SIV community.

BB

Bruce Balentine
EVP and Chief Scientist, EIG Labs
517 Roberts
Denton, Texas 76209
940-891-3414 (office)
940-206-9524 (mobile)
940-382-2143 (home)
http://www.eiginc.com

# # #