



Proposal for Presentation of Discussion Poster

Important Open Questions: Telephone-based Dialog Systems

Lin L. Chase
Speechworks International, Inc.
Lin.Chase@Speechworks.com

Per a discussion with Paul Heisterkamp in Cambridge on January 21, 1999, I would like to suggest the following poster topic for you "tea lounge discussion area". This is simply a list of open questions that I find important in dealing with the development and deployment of live spoken language dialog systems, as situated in actual industrial contexts. This submission is certainly not a scientific contribution in any sense other than it poses some questions that may possibly interest researchers. I would of course be excited to find that some of them have already been answered

Evolution of Design of Call Flows and Interactive Styles

Many currently deployed telephone-channel dialog systems are built around technologies that have the following characteristics

- Isolated word (or very short phrase) speech recognition
- Speaker-dependent speech recognition
- No barge-in supported, beeps played after prompts to signal system readiness, sometimes "you spoke too soon" beeps played back to speaker
- Rejection used, but not confidence measures; this leads to use of static strategies for confirmation in call flows.

We now have the ability to deploy telephony systems that use:

- Large vocabulary continuous speech recognition
- Speaker-independent speech recognition
- Barge-in supported, beeps after prompts optional
- Confidence measures used together with dynamic call flow topologies

This should be good news to the consumers of our technology! It's not always the case, however, that clients with older technology deployed are overjoyed to hear about these developments. For a variety of political and user acceptance reasons, we are often faced with the need to provide a very smooth, slow, and practically invisible shift in functionality from the former to the latter.

How can this gentle evolution be accomplished? What is the right sequence of changes to introduce into a deployed system? At what pace?

Cultural Issues in Prompting Strategies

We are experienced in deploying telephone-based systems in the context of North American cultures, but much less in Europe, Asia, and Central and South America. Initial work in new contexts has led us to understand that our current guidelines for developing a “personality” for a system may be somewhat culturally bound. For example, expert critics from Europe have found our interfaces to be over-friendly and overly apologetic.

What needs to change in order to make a system “fit” in the various contexts outside of North America?

- How direct or personal should the prompts be?
- How obviously instructional or directive (as opposed to supportive and suggesting)?
- How friendly (as opposed to serious-sounding)?
- How apologetic when errors are detected or suspected?
- Does the order in which information is collected change when asking for basic kinds of information such as addresses, phone numbers, account numbers?
- Do turn boundaries always correspond to single transactions, or are multiple turns sometimes used for a single act?
- What linguistic politeness forms are appropriate
- Are there major differences between how the system should interact with novice users and experts?
- Are there other major axes that should be considered?

Measuring What Really Works for Callers

As a community, we don't have a good working theory of interaction between human callers and computers that listen and talk. We also don't have much experimental data about what makes a dialog system usable by its callers, nor about what makes one system better than another.

Unsupported claims about issues such as the appropriateness of confirmation, the usefulness of barge-in, and the relative importance of recognizer performance compared to transaction completion rate. While these claims may in fact be dearly held beliefs based on the individual experiences of the claimants, they are unfortunately not backed up by any kind of general or controlled experimental work, and are therefore suspect.

What work exists already in measuring the following? What can we do to create an experimental approach for learning more?

1. Analyze the effects of
 - TTS vs. recorded prompts and playback
 - Use of mixed TTS and recordings (where TTS is used for dynamic vocabulary additions, for example)
 - Use of barge-in at all
 - Use of simple barge-in vs. barge-in based on acceptance of an understood utterance
 - Use of confirmation at all
 - Use of confirmation under suspected error vs. "worst case only" situations
 - Use of confirmation for sensitive information only (such as currency amounts)
 - Use of don't-repeat-mistake strategies in call flow
 - Use of highly directed vs. open-ended prompts
 - Different call flow designs for novice and expert users vs. strategies that work for both
 - Insertion of promotions and newsworthy announcements, including their placement in the call flow

2. Use the following measures
 - Recognition word error
 - Recognition utterance error
 - Turn error (utterance-level errors not related to recognizer performance)
 - Transaction completion rate
 - Duration of call
 - Percentage of callers who simply give up
 - Caller subjective measures: annoyance at having fallen upon a computer, general level of frustration with system, repeat visit rate when human callers are available as an alternative, ...

What other system characteristics should be studied? What other measures should be considered? Which measures should be used on which characteristics? How should such studies be performed? Should callers be aware of measurements during calls? What can be measured if they're not? Can we contact real users later somehow? If we do, will they be able to give us useful feedback?

