

Evaluating real callers' reactions to Open and Directed Strategy prompts

Jason Williams, Andrew Shaw, Larry Piano, Mike Abt

Edify Corporation
2840 San Tomas Expressway
Santa Clara, CA 95051

+1-800-713-3439 or +1-408-982-2030

{jason.williams, andrew.shaw, lawrence.piano, michael.abt}@edify.com

Abstract

Real callers' first reactions to an Open or Directed Strategy prompt for a call routing application in the consumer retail industry are analysed. Each prompt is trialled alone, with an earcon, and with an earcon and named persona in a total of six experiments. Usage of the earcon & named persona is found to maximize "routability" in all cases. The Directed strategy achieved higher "routability" and lower rates of user confusion than the Open Strategy for all experiments, implying that the Directed strategy may be better suited for this caller base. Application of other prompt wording styles and analysis of follow-on behavior are promising areas for future study.

Introduction and motivation

"How may I help you?"-style prompts are increasingly being used in commercial speech-recognition applications for call routing. This type of *Open Strategy* prompt is often preferred to a *Directed Strategy* prompt because it is believed to increase caller satisfaction, task completion times, and task completion rates.

This paper seeks to compare real callers' first reactions to an "Open" strategy vs a "Directed" Strategy prompt by assessing task completion and measures of confusion for each.

Further, naming the system (i.e., the automated system introducing itself with its name) and using earcons are increasingly being used in commercial systems. This paper seeks to determine what effect these strategies have on callers' very first reactions for both the Open and Directed strategies.

Several studies have attempted to assess how prompt strategy affects user satisfaction and task completion rates, but none have applied this specifically to the call routing task.

[1], [2], and [3] give general insights into task completion and caller preference for open vs. directed prompts, but are not applied to the call routing task. [1] and [2] tested a train timetable system of various levels of "open-ness". [1] found that user satisfaction is shown to derive from user perception of task completion, recognition rates, and amounts of barge-in. "No efficiency measures are significant predictors of performance" for real systems, but "in the presence of perfect ASR, efficiency becomes important." [2] found that "Strategies that produce longer tasks but fewer misrecognitions and subsequent corrections are preferred by users." [3] explored a Yellow-pages search task. The study found that explicitly listing fewer choices using more turns was found to be more preferred by users than asking open questions among more choices using fewer turns. "It is interesting to note that even in the case of a search task, where the interaction itself should not matter as much as the information to be retrieved, users did not necessarily prefer the interface that would take them the fastest to the desired result."

[4] and [6] studied user behavior in open prompts used for call routing. [4] compared a variety of wordings for an "Open" prompt to find an approach that maximized task completion and caller preference in a call routing task, but did not compare it with a directed approach. [6] assess responses to a human operator's "How can I direct your call?"

[7] studied usability subjects' responses to a banking application's first prompt – a call routing task. Three prompts were compared:

Style	Wording	% Silent	% Containing a keyword in first try
<i>Open</i>	"Main menu – how can I help you?"	6%	15%
<i>Mid</i>	"Main menu – which service do you require?"	16%	48%
<i>Closed</i>	"Main menu – Please say 'help' or the name of the service you require."	10%	44%

Users reported significantly higher rates for "knowing how to select options" (from Likert scores) for the *Closed* style than the others. However, overall, no clear preference between the systems was reported. 69% of subjects asked for "help" in the first turn in the *Closed style* experiment.

[7] differs from the present study in three ways: (1) it studied the reactions of usability subjects rather than real callers, (2) "routability" was assessed assuming keyword spotting rather than assessing information content and assuming an SLM, and most importantly, (3) none of the menu choices explicitly listed choices, whereas the results from [7] demonstrate that when callers know how to get a list of choices, the vast majority will request them. The present experiment lists the routing choices.

Background

The project was initiated for an Edify customer in the US consumer electronics industry. In this paper we will refer to the company as Acme.

In this project, Acme was seeking to route calls using ASR based on *task area*. Each task area and estimated % of calls (as determined by today's DTMF call routing system) are given below:

Task area	Description	% of calls
Technical support	Phone-based product support – e.g., product usage questions, hook-up explanations, how-to and step-by-step guides	61%
Product information	Information about products or accessories that a caller is considering purchasing	14%
New repairs	Both finding a local repair store, and arranging to mail the product to Acme to be repaired.	10%
Store locations	Finding a location of a nearby store where a caller can buy a particular product	8%
Existing repairs	For repairs that Acme is currently performing, playing back the status of the repair – e.g., "A technician has examined it and we expect to be sending it out in 7 working days."	7%

Most task areas also required capturing the product category (for example, televisions) and/or model number (for example, model XYZ-123) to successfully route the call. Acme sells thousands of different products.

For each task area & product category pair, Acme sought to automate a sub-set of caller tasks, such as sending a fax with product specifications, playing back location of stores where a caller could find a particular product, or providing information about getting a product repaired.

Where automation was either impractical, or where automation was attempted and failed, callers are transferred to agent queues based on both task area and product category.

Assumptions

It was contemplated whether it would be possible to reliably recognize in one utterance (or reliably determine the absence of) product category *and* task area. We assumed that it would not be possible to extract this much information from an utterance reliably enough for a commercially-deployed system. Further, we reasoned that confirmations and error handling would be cumbersome, and frequent.

Thus, we assumed that the routing problem can be approached by identifying the *task area and product category/model number* for a given call **where each must given in a separate utterance.**

Since not all task areas require a product (e.g., calling to find the status of a repair), it was natural to make the first interaction focus on task area. This paper examines several alternatives for this first interaction of the call.

Methodology

Goal

We identified three possible introductory greetings:

Greeting	Prompt text
Hello	“Hello, welcome to Acme”
Hello + Earcon	<i>[earcon]</i> “Hello, welcome to Acme”
Hello + Earcon + Persona & recently changed notice	<i>[earcon]</i> “Hello, welcome to Acme. My name is Johnson, your virtual assistant. Please take note, this service has recently changed.”

We also identified two initial prompts to reflect the routing strategies:

Routing Strategy	Prompt text
Directed	“Please tell me which of the following options you’d like: tech support, repairs, product information, or store locations.”
Open	“What can I help you with? <i>[2.5 sec pause]</i> You can get tech support, product information, repair information, or store locations. Which would you like?”

These were combined to create 6 experiments:

Experiment	Greeting	Prompt
1	Hello	Directed
2	Hello	Open
3	Hello + Earcon	Directed
4	Hello + Earcon	Open
5	Hello + Earcon + Persona & recently changed notice	Directed
6	Hello + Earcon + Persona & recently changed notice	Open

Our primary goals were to understand, for the very first interaction of the call:

- Which routing strategy had the highest perfect-ASR task completion rate?
- What were reasons for failures for either strategy?
- Does adding an earcon alone improve perfect-ASR task completion rates? Adding an earcon, persona, and change notice?

System design

Each of the strategies above was implemented using a Wizard-of-Oz (WoZ) based system. An operator (or “wizard”) selected the system’s response from a pre-defined set of options in each interaction (e.g., rejection, selection of next state, transfer to an operator, etc.). The telephony interface supported barge-in throughout all prompts, and the speech/no-speech decision (i.e., activation of the end-pointer indicating user speech) was made by the telephony card (not the wizard). The same voice talent was used for all, and the voice coaching & persona attempted to maintain consistency.

Each interaction consisted of playing the initial prompt, with barge-in, and waiting for a user response. The first utterance (or lack thereof) was collected.¹

Data collection

A subset of incoming calls from Acme’s main (national) toll-free phone numbers were diverted to a group of agents trained on the WoZ system. After interacting with the WoZ simulation, callers were transferred to agents.

Calls were taken during business hours in morning and afternoon shifts.

Given the small fraction of calls taken by the system, and the relatively low number of repeat callers among Acme’s caller base, we believed it was unlikely that more than a fraction of one percent of the calls taken in this data collection had called a previous data collection.

¹ To support other research, each experiment also had 2 escalating no-input (i.e., silence) reprompts which were played when no speech was detected. Each interaction also had 2 escalating no-match (i.e., invalid/out-of-grammar) reprompts. The interaction continued to also ask for product category and/or model number. If a subject triggered more than 2 no-inputs or more than 2 no-matches in a given interaction, an agent transfer was simulated.

Utterance classification

The caller’s first response to each interaction was classified into one of the following categories:

Category		Sub-category label	Sub-category description
<i>Invalid (excluded from analysis)</i>		I	Invalid utterance – background noise, voices, end-pointer error, etc. <i>Note that invalid utterances have been excluded from the percentages calculated for each of the categories</i>
<i>Routable (success)</i>		R	Routable: utterance contained enough information (perhaps over-informative – e.g., included product category) to route the call to one of the task areas
<i>Failures</i>	Confusion (Conf)	S	Silence – end-pointer didn’t trigger (typically indicating the caller didn’t speak) ²
		C	Obvious caller confusion – e.g., “Hello?” “Is this Acme?” “Is this a real person?”
	Non-cooperative (non-coop)	H	Hang-up
		D	The caller pressed a DTMF key (note that the prompts did not include any instructions for DTMF)
		A	The caller explicitly requested to speak with an agent/real person
	Content (Cont)	U	Cooperative caller but unroutable for any reason not listed above (e.g., the caller said just a product name, or gave insufficient information to determine a task area, such as “Yes, hello, I have a question.”)

Results

Exp	N (total)	I	N (valid)
1	266	0	266
2	64	1	63
3	209	7	202
4	56	0	56
5	152	8	144
6	52	5	47

Exp	N (valid)	R	Conf			Non-coop			Cont
			S	C	H	D	A	U	
1	266	44%	42%	3%	6%	0%	0%	4%	
2	63	29%	37%	24%	2%	0%	3%	5%	
3	202	52%	37%	4%	3%	2%	0%	0%	
4	56	34%	13%	34%	5%	0%	0%	14%	
5	144	61%	26%	2%	8%	1%	1%	1%	
6	47	43%	21%	17%	6%	0%	0%	12%	

The primary reason for *U* (Failure due to utterance content) was callers saying the name of a product in isolation with no indication of task – such as “Television.” In experiment 4, 10% of valid utterances contained a product in isolation (and 4% contained no task or product

² Here we classify silence as a type of confusion – i.e., the caller didn’t know what to say, or didn’t know that they should speak. “It has been observed that if the caller experiences problems with the service or becomes confused about what response is expected, then they will tend to remain silent while they work out what to do.” [5].

information, like “I have a question.”). In experiment 6, all 12% of *U* utterances are products in isolation.

Figure 1 shows percentage of first utterances which were “routable.”

Figure 1: *R* vs. Introduction type for Directed and Open strategies

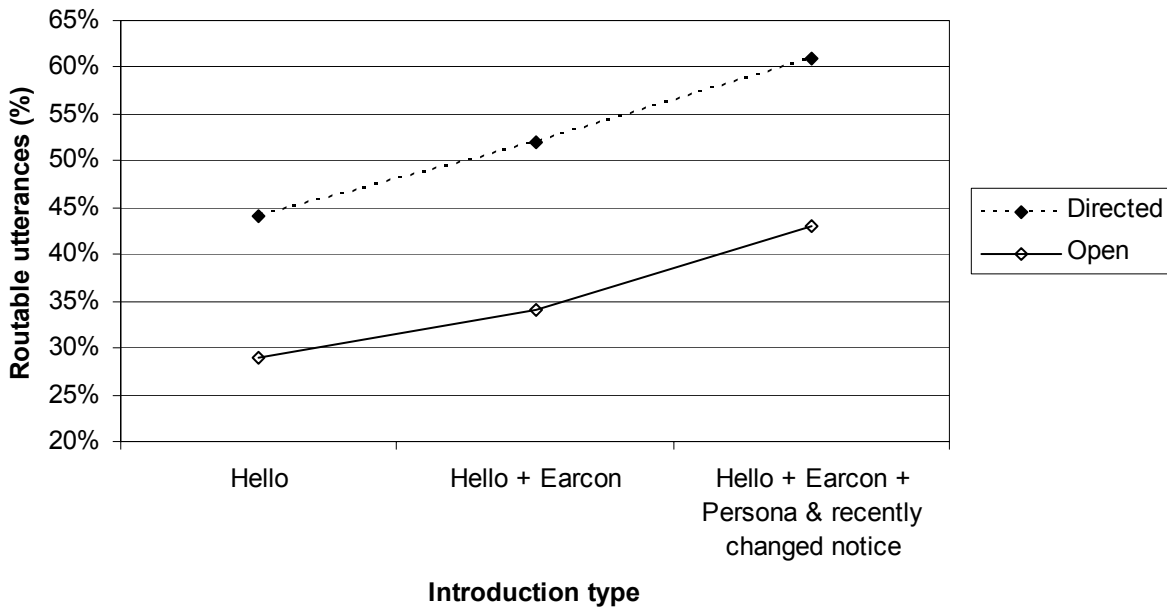
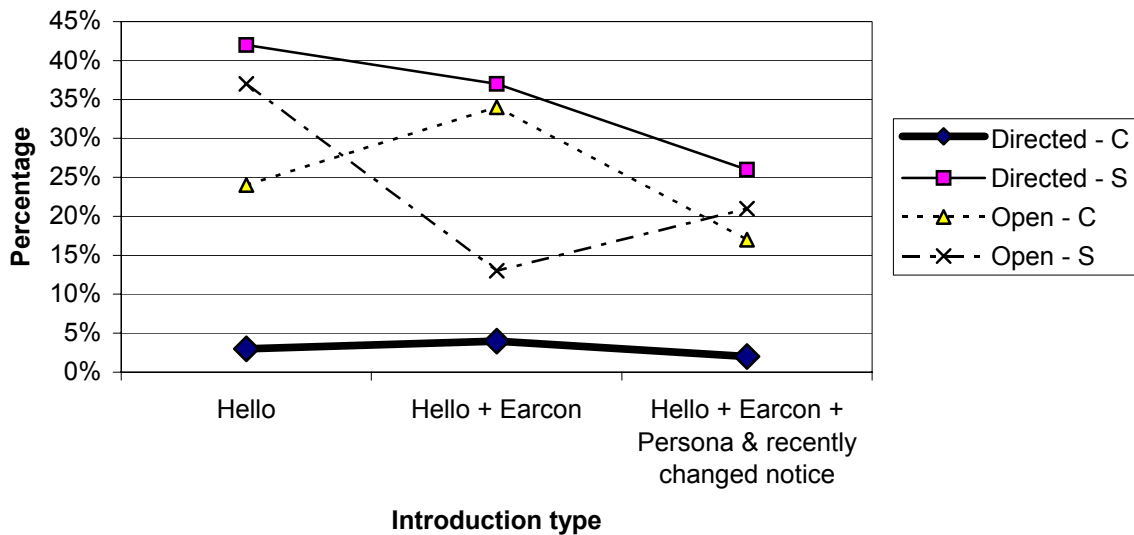


Figure 2 shows reasons for confusions (the primary reasons for failures) – *C* (clear user confusion) and *S* (user silence) in each experiment.

Figure 2: Primary failure reasons



Figures 3 and 4 show the following results for the Directed and Open strategies, respectively: *R* (Routable), *Conf* (all confusion types – silence or verbalized confusion), *Non-Coop* (non-cooperative user), and *Cont* (unroutable due to utterance content).

Figure 3: Categorization vs. Experiment (Directed strategy)

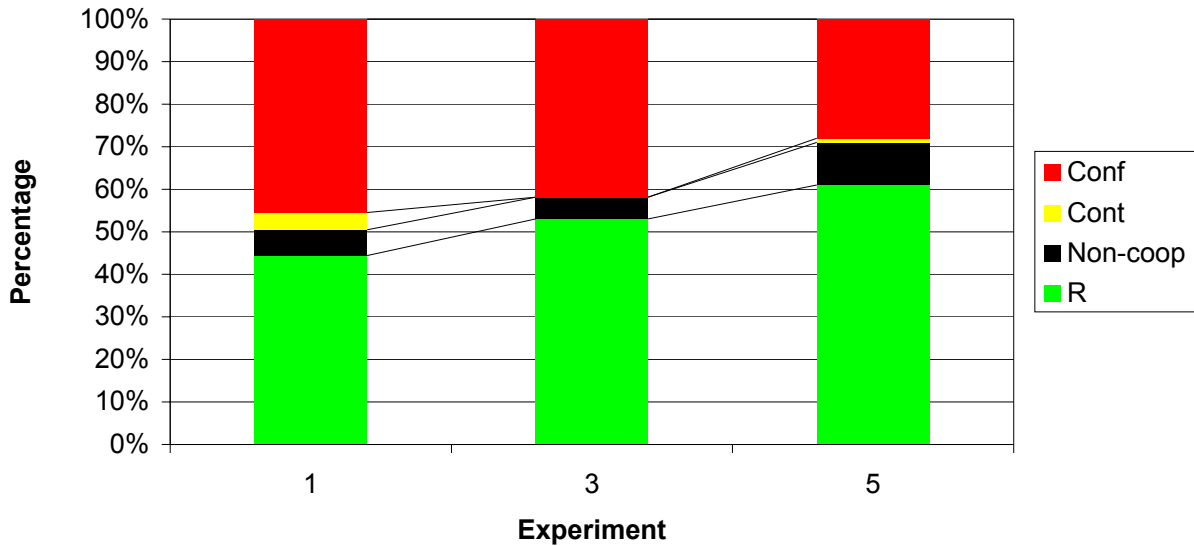
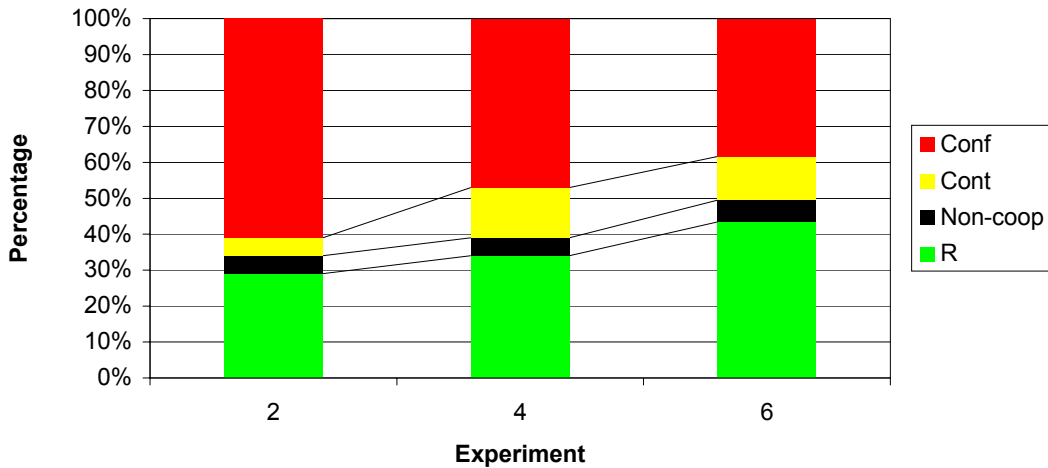


Figure 4: Categorization vs. Experiment (Open strategy)



Conclusions & future work

We hypothesized that *U* (unroutable due to content) utterances that contained only a product but no indication of task could be attributed to the language of the prompt – “what can I help you *with*?” as opposed to “What would you like to do?” or “How can I help you?”. If we count utterances with product in isolation as members of *R* (routable), this diminishes the gap in

routability between the Open and Directed strategies, but does not alter the difference between the two.

One anomaly was the increase in silence from experiment 4 to 6 (see figure 2, up-trend of line marked “Open-S”). We noted that this was accompanied by a commensurately larger decrease in verbalized confusion. We concluded that adding the named persona and change-notice from experiment 4 to 6 had the effect of trading some confusions for silences, and others for routable utterances. We note that the sum of verbalized confusions and silences decreased from experiment 4 to 6 (see Figure 4).

We concluded that the directed strategy was more likely to succeed on the first try for this domain. Further, we concluded that the Open strategy (without further introductory prompting, which was not explored) caused approximately an order of magnitude greater level of confusion among callers in the first utterance.

We had expected that *U* utterances would form a larger portion of Open strategy responses. This study implies that user confusion/motivation (as expressed by silences, hang-ups, and verbal expressions of confusion) is a larger problem at the first prompt than information content.

Other research (such as [4]) has shown that the phrasing of an Open strategy prompt can have a significant impact on utterance content – particularly, the choice of example phrases (keywords vs. natural-language phrases) and the placement (i.e., before or after the question.) This would be an appropriate area for future study.

In addition, other work we conducted in usability tests suggested that callers prefer (in this domain) an interactive series of questions above both the Open and Directed Strategies. Testing this approach on live callers is another area of further promising study.

References

- [1] Diane J. Litman, Shimei Pan, and Marilyn A. Walker. “Evaluating Response Strategies in a Web-Based Spoken Dialogue Agent.” In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and the 17th International Conference on Computational Linguistics (COLING-ACL'98)*, pp. 780-786, Montreal, Canada, August 1998.
- [2] M. Swerts, D. Litman, & J. Hirschberg. “Corrections in spoken dialogue systems.” *Proceedings of the International Conference on Spoken Language Processing (ICSLP 2000)*, pp. 615-618, Volume II, Beijing, China, 2000.
- [3] W. Vincent Vanhoucke, Lawrence Neeley, Maria Mortati, Michael J. Sloan and Clifford Nass. “Effects of Prompt Style when Navigating through Structured Data” (with accompanying presentation Speech and Search). *Proceedings of INTERACT 2001, Eighth IFIP TC.13 Conference on Human Computer Interaction* (IOS Press), pp.530-536, Tokyo, Japan, 2001.
- [4] Tony Sheeder. “Learning From User Performance: A Laboratory Study.” *Nuance V-World 2002* (Nuance), Orlando, Florida, 2002.

[5] F. Stentiford, P. Popay. "The design and evaluation of dialogues for interactive voice response services" *BT Technical Journal* (British Telecom: United Kingdom). No 1, Vol 7, January 1999. pp 160-171. pp 142-148.

[6] B. Carpenter, J. Chu-Carroll. "Natural Language Call Routing: A Robust Self-Organizing Approach." *Proceedings of ICSLP 1998*, Sydney, 1998.

[7] F. McInnes, I. Narin, D. Attwater, M. Jack. "Effects of prompt style on user responses to an automated banking service using word spotting." *BT Technical Journal* (British Telecom: United Kingdom). No 1, Vol 7, January 1999. pp 160-171.