EVALUATING SPOKEN DIALOG SYSTEMS FOR TELECOMMUNICATION SERVICES

Candace Kamm, Shrikanth Narayanan, Dawn Dutton, Russell Ritenour

AT&T Labs–Research
180 Park Avenue, Florham Park, NJ 07932
cak@research.att.com shri@research.att.com
dldutton@att.com rit@research.att.com

ABSTRACT

This paper presents a case study analyzing the results of an on-going trial of a prototype mixed-initiative spoken dialog system for telephony control and messaging. System usage and performance data were captured at three points in time. Information from multiple data sources, including spoken utterances, system call logs, speech recognizer output, and subjective surveys was evaluated to determine the relationship between aspects of system performance and user perceptions of the system. This report provides several examples using these data sources in combination to identify key areas to focus on in modifying the system, application, and/or user interface in order to significantly improve system usability and user satisfaction.

1. INTRODUCTION

Spoken dialog agents that enable easy and friendly automation of services such as communication assistance and business transactions are gaining increasing popularity[1, 2]. One of the major issues in the development of these systems is how to evaluate both overall performance of the system and performance of the component technologies[3]. We have developed a prototyping platform offering a variety of telephone-based services. The system is currently being used by 35 "friendly" users - employees and their families. This platform has allowed us to collect data from a variety of sources (e.g., speech utterances, speech recognition results, logs of user interactions with the system, user perception and satisfaction surveys). In this paper, we present a case study of how these data sources are used to evaluate ongoing system performance, to understand the relationship between system performance and user perceptions of the system, and to identify key aspects of the system that require modification to improve system usability and user satisfaction. Our goal is to take a user-centric view of the system, analyzing system performance in terms of the users beliefs about how the system should perform, in addition to the more traditional system-centered, technology driven evaluation techniques.

1.1. SYSTEM DESCRIPTION

Our prototype system, Annie, is a multi-channel, client-server system that supports a mixed-initiative voice-enabled application for call control and messaging. The current functionality includes dialing from a personal directory, dialing by spoken numbers, access to employee directories, voice-based and web-based label administration (for entering, deleting and modifying the entries in the user's personal directory), and message creation and retrieval. The underlying speech recognition technology uses context-dependent phone models for telephone speech, with constrained grammars defining the legal vocabulary for each feature. The system supports barge-in and DTMF (touch-tone) defaults for some functions. The user interface to the system uses an anthropomorphic "personal assistant" metaphor to achieve a "conversation-like" interaction, within the limitations of the constrained-grammar ASR technology.

The system is available 24 hours a day, and users use it both from home and away from home. Each user's home and office number are known to the system, so when the user calls in from either location, the system automatically detects the incoming number and retrieves the user's personal directory. Access from other locations requires the user to speak an account number or to enter it using touch tones. The system handles an average of 120 calls per day, from an average of 20 different accounts. There is no human attendant or other customer care in case of system failures.

For each interaction with the system, the following information is automatically recorded: a) each spoken utterance that the recognizer processes; b) the ASR output corresponding to each spoken utterance, including the recognized word, recognition scores, rejection scores, barge-in status, endpoint information, vocabulary and grammar information; and c) a log of the sequence of prompts, recognitions, database access, and call detail events that occur during the interaction.

2. TESTING METHODOLOGY

2.1. SUBJECTS

The 35 "subscribers" to the Annie service are all employees who have experience with voice-enabled technology. Because subscribers were encouraged to have members of their families also use the service, some accounts had more than one person using the system, and the user population had varying degrees of previous experience with speech technology. About two-thirds of the subscribers were native speakers of English. As an incentive to use the system, calls placed through the system to locations in the U.S. and Canada were free to the subscriber. International calls and calls to toll-free (800, 888) or 900 numbers were not permitted.

2.2. DESIGN AND PROCEDURE

To study the evolution of system performance and users reactions to the system, data were collected during three time periods. In the first period, users were asked to perform a set of scripted scenarios on each of five days to test specific features of the system. Because many of the subscribers were new to the system at that time, the scenarios served as structured "training" sessions, introducing the users to the system features. An example scenario for repertory dialing features is shown in Figure 1. Similar scenarios were used to test dialing by speaking the phone number, adding and deleting voice labels, and accessing the employee directory. For each utterance they spoke during the scenarios,

Dial 1-800-xxx-xxxx.
While the "Account Number Please" prompt is being played,
say *<your account number>*.
While the "Hello, Annie here" prompt is being played,
say *Call Jay Wilpon*.
After you start hearing ringing,
say *Cancel*.
While hearing "I cancelled it. Annie here"
say *List my labels*.
While listening to the labels, interrupt playing and
say *Goodbye Annie*.

Figure 1: Sample User Scenario

users specified on the scenario answer forms whether the system recognized their input speech correctly, and if so, how many attempts it took the system to do so. At the end of the five days of scenarios tests, users were given a survey

that asked for ratings on their perceptions of overall system performance and whether the system features worked as expected.

Following the initial scenario testing, users were encouraged to use the system for their everyday telephony needs. A followup user survey was performed three months after the scenario tests. This survey asked questions about usage of and satisfaction with different features of the system. To associate objective measures of system performance with the survey data, the corresponding speech utterances and system logs for all calls made during the week prior to the survey were also captured. A third "snapshot" of speech data and system logs was collected three months later.

3. RESULTS

3.1. USER SURVEYS

During the scenario test period, there were several significant system problems that influenced user reactions and perceptions. For example, 73% of the 22 respondents to the initial user survey experienced failures where the system did not respond for a prolonged period, and half of the users also experienced the system terminating the interaction unexpectedly. The system was described as being too slow by 73% of the users, and 27% of the users indicated that they had trouble interrupting system prompts and canceling actions. Despite these factors, only 2 users (9.5%) found it somewhat difficult or difficult to use.

On the follow-up survey, 23 of 28 respondents (82%) reported having used the system during the previous 4 weeks. Lack of system responsiveness was experienced less frequently than in the initial scenario tests. This improvement was the result of modifications in the user interface to detect system error conditions, correct them when possible, and alert users appropriately otherwise. Of those who used the system, 78% reported that it was extremely easy or moderately easy to use, and only 1 person (4%) found it somewhat difficult to use. Over 80% of users reported that they were "almost always" successful using the following features: accessing the system via a spoken account number, accessing help messages, listing personal labels, and using the web page for label administration. Overall, 87% of users were "very satisfied" or "satisfied" with the recognition performance for voice labels, 70% with ASR for commands, and 74% with ASR for digits. Fewer users (32%) expressed dissatisfaction with the speed of the system than during the scenarios tests. Subjective comments primarily addressed slow system response, variable recognition performance, and lack of robustness and reliability of the system.

3.2. SPEECH RECOGNITION PERFORMANCE

To determine whether objective measures of speech recognition performance correlated with the user perceptions, the speech utterances for each account collected during the scenario period and each usage period were transcribed by human listeners. The transcribers classified the speech files into the categories shown in Figure 2, and also marked whether any unusual non-speech events occurred during the utterances (e.g., background noise or speech, breath noise).

The tree structure shown in Figure 2 was used to classify user input for analysis. Utterance files were classified as containing speech or no speech. Files containing speech were further broken into four categories: utterances that contained only legal words or phrases in the recognizer grammar (in-vocabulary), utterances that contained legal words and phrases, but embedded in other extraneous speech (e.g.,"Call home, please" utterances that consisted of multiple repetitions of legal words and phrases (e.g., "Cancel. Cancel.") and utterances that consisted only of out-of-vocabulary (OOV) speech. The OOV speech was further separated into four sub-groups: utterances that were like a voice label, but not in the vocabulary, utterances that were similar to a command, but not in the vocabulary, utterances that were related to a label or command in the vocabulary (e.g., saying "Call Candy Kamm" when the voice label in the user's vocabulary was "Candace Kamm" and utterances that consisted of irrelevant speech. Table 1 presents the distribution of utterances spoken when each grammar was active into the five input categories shown in Figure 2. The first three grammars are identical for all users. The "personal grammars" consist of all grammars that included the user's personal dialing list (e.g., the grammars used at the top level of the system, in out-bound calling, in voice label administration, and in messaging). Across all accounts,

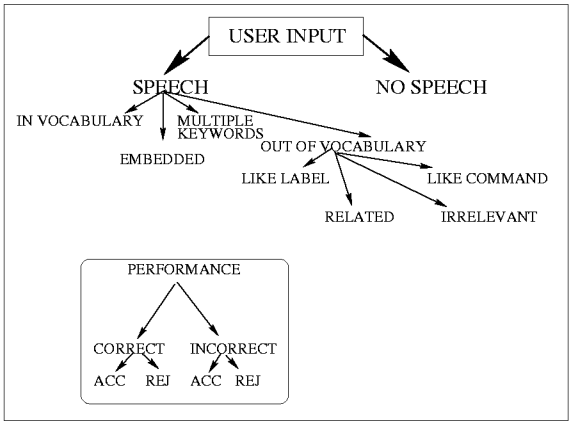

Figure 2: Classification of speech data categories. Inset: recognition performance categories (acc - accepted, rej - rejected).

Grammar	# Utt.	Test Period	In Voc.	Emb.	Mult.	OOV	No Speech
Cancel	1138	A	52.8	2.1	8.7	1.5	34.9
at	979	B	12.6	1.0	3.4	3.8	79.2
Dialing	897	C	14.2	0.8	3.0	1.5	80.5
Employee	840	A	49.4	4.6	1.3	26.5	18.1
Directory	46	B	76.1	0.0	0.0	15.2	8.7
	96	C	45.1	0.0	0.0	36.6	18.3
	1296	A	68.7	3.3	0.2	11.4	16.4
ID Number	427	B	55.0	1.6	0.0	5.1	38.1
	356	C	59.5	0.6	0.0	3.9	36.5
Personal	3659	A	43.3	9.4	6.0	12.9	28.2
Grammars	1497	B	77.9	5.5	0.9	7.8	7.8
	1536	C	73.7	4.9	1.7	8.1	10.5

Table 1: Total number of utterances (# Utt.) and proportion of utterances in each input category for three test periods. Test periods are scenario testing (A), usage 1 (B), and usage 2 (C).

personal grammars were active for 54-62% of the utterances during all three time periods. In contrast, the employee directory feature and the remote access ID number feature were accessed relatively less frequently during the usage periods than during the scenarios when users were required to exercise those features. Table 1 shows that the proportion of in-vocabulary utterances decreased significantly for the "cancel at dialing" grammar in the usage periods as compared with the scenario tests. This was because users were instructed to cancel out-bound calls during the scenario tests, but in normal usage, most calls were placed, so utterances captured during the time the "cancel at dialing" grammar was active typically contained no speech. The relative increase in in-vocabulary utterances for the "personal grammars" in the usage periods may reflect users becoming more familiar with the system vocabulary and dialog pacing. The increase in OOV utterances for the employee directory grammar in usage period 2 was due to the fact that the directory had not been updated and users were requesting employees who should have been in the directory but had not yet been added to the recognition grammar. The relative increase in utterances containing no speech (and decrease in in-vocabulary utterances) for the ID number grammar between the scenario period and the usage periods reflects the more frequent use of DTMF touch-tones to enter the account number during the usage periods.

The speech recognition result for each utterance was compared with the utterance transcription to determine whether recognition was correct or incorrect (see inset, Figure 2). The recognizer used a rejection criterion, so each recognition result could either be accepted or rejected. Thus, for each type of speech input, the recognition outcomes were classified into one of four bins: a) correctly recognized utterances, b) incorrectly recognized utterances, c) utterances that were correctly recognized but were rejected, and d) utterances that were incorrectly recognized and (correctly) rejected. This categorization was used to determine recognition "success" rates using four different metrics, defined in Table 2.

Metric	Numerator	Denominator
HC_{U1}	total in-vocabulary correctly recognized	total in-vocabulary utterances
HC_{S1}	total in-vocabulary correctly recognized or correctly rejected	total in-vocabulary utterances
HC_{U2}	total utterances correctly recognized	total utterances with foreground speech
HC_{S2}	total in-vocabulary, embedded and related utterances correctly recognized or correctly rejected	total utterances with foreground speech

Table 2: Recognition "Success" Metrics

Grammar	Test Period	HC_{U1}	HC_{S1}
Cancel at Dialing	A	92.51	95.34
	B	87.10	97.58
	C	84.25	93.70
Employee Directory	A	79.67	90.91
	B	78.37	84.55
	C	67.99	84.57
ID Number	A	70.34	79.55
	B	74.04	81.28
	C	81.43	87.14
Personal Grammars	A	63.67	81.82
	B	72.50	81.57
	C	73.44	81.30

Table 3: Speech recognition success rates (%) for four grammars, using the first two metrics defined in Table 2. Test periods are scenario testing (A), usage 1 (B) and usage 2 (C).

Metric HC_{S1} considers how well the recognizer performs, given that it is presented with utterances consisting only of legal vocabulary. Both correct recognition and correct rejections are considered "successful" outcomes, because the system has not committed an overt error in either case. From a user-centric viewpoint, however, any recognition outcome to a legal input utterance that does not result in progress toward task completion might be considered a negative outcome. Metric HC_{U1} takes this view, and considers only correctly recognized in-vocabulary utterances in its numerator. Metric HC_{U2} takes an even more radical view of "success" considering only correctly recognized utterances as "successful" but expanding the set of utterances to include all speech files that contained foreground speech (i.e., speech by the subscriber). Metric HC_{S2}, which considers both correct recognition and correct rejection as "successful" outcomes, is the system-centric analog to HC_{U2}.

A summary of the speech recognition results from utterances collected during the scripted scenario tests (A) and the two regular usage periods (B and C) is presented in Table 3. Mean results are presented for the HC_{U1} and HC_{S1} metrics and show the recognizer "success" rates for four of the grammars used in the application. The results shown in Table 3 for "personal grammars" are averages over the 25 accounts that used the system during all three collection periods. The two metrics shown in Table 3 show slightly different trends across grammars and test periods. [1] For the "personal grammars", HC_{S1}, the system-centric measure considering only in-vocabulary utterances, does not change appreciably across test periods, but HC_{U1} shows an improvement from the scenario period to the usage periods. In contrast, the "ID number" grammar shows an improvement across time for both metrics. These differences may be explained by changes that occurred in both the system and usage patterns across test periods. Plausible contributors to performance improvements for both grammars include a) increasing the system's time-out parameters, b) eliminating a system board that was distorting the incoming speech, c) users learning how to use and adapting to the system, and d) attrition of users who had difficulty with the system. Eliminating a system problem with voice label creation that corrupted some personal grammars during the scenarios tests no doubt contributed to the improvement in success rates for the personal grammars, as did adding a facility for automatically generating voice label pronunciations from text.

[1] The four metrics described in Table 2 were all significantly correlated, with HC_{U2} and HC_{S2} following the same general trends as HC_{U1}.

Grammar	Test Period	Complete on 1 Attempt	Complete on 1 or 2 Attempts	Not Completed
ID Number	A	56.3	73.6	13.7
	B	70.1	84.9	10.2
	C	74.7	89.7	7.8
Personalized Grammars	A	52.1	73.4	11.3
	B	79.9	91.2	3.6
	C	81.2	94.3	1.5

Table 4: Proportion of ID Number and Out-bound Calling Tasks successfully completed on first attempt or first or second attempt (%) for three test periods derived from System Log Data. Test periods are scenario testing (A), usage 1 (B), and usage 2 (C).

Grammar	Test Period	Completed on 1 Attempt	Completed on 1 or 2 Attempts
ID Number	A	68.2	86.5
	B	64.0	100.0
Personalized Grammars	A	56.8	82.0
	B	57.0	90.0

Table 5: Proportion of ID Number and Out-bound Calling Tasks successfully completed on first attempt or first or second attempt (%) for three test periods from Subjective User Estimates. Test periods are scenario testing (A), usage 1 (B) and usage 2 (C).

Correlations were computed between the user satisfaction ratings for ASR performance for voice labels, digits, and commands and the metrics for the subset of 18 subscribers who both used the system during all three test periods and completed the survey. None of the correlations reached statistical significance. HC_{U1} had the highest correlations with user satisfaction ratings, with correlations of 0.41 between user satisfaction ratings for personal voice labels and ASR performance for voice labels, 0.44 between user satisfaction and ASR performance for commands, and 0.25 between user satisfaction and ASR performance for digits. One possible explanation for the weak correlations is that user ratings may reflect a holistic view of system performance rather than its components, and users may not be able to provide judgments about any specific component (e.g., ASR performance or reliability) that are independent of their overall experience with the system. A second contributing factor is that users who were truly dissatisfied with the system may have stopped using it entirely, and so they would not be contributing data to the correlational analysis.

4. SYSTEM CALL LOGS

To provide a task-oriented view of system performance, the system logs for each call during the scenario and usage periods were analyzed. During the scenario tests, about two-thirds of the calls to the system were made to the remote access number, which invoked the ID Number grammar for establishing the subscribers identity. During the usage periods, only about one-third of the calls used the remote access number to get to the system.

Table 4 presents an analysis of the number of attempts required to complete the ID number task and the out-bound calling task. The results in this table demonstrate that task completion rates on both one and one or two attempts increased across the three test periods. Table 5 shows perceived task completion rates, as estimated by the users on the surveys taken during the scenario tests and after the first usage period. On average, the users overestimated task completion rates on the scenario tests and underestimated task completion rates for a single attempt during the usage period.

5. RELATING MULTIPLE SOURCES OF INFORMATION

One of the major impediments to a controlled evaluation of the factors influencing performance of a system over the time is the fact the system generally evolves in many ways, so it is difficult to attribute improvements to any single

factor. For example, in our system, the scenario tests themselves uncovered several major problems with the quality of the speech coming into the system, system reliability and responsiveness, and system parameter settings such as ASR time-out parameters. Because the system was being used as a "real" service, such problems were fixed as soon as possible. In addition, because the system was also being used to test new features, the feature set and command grammars expanded over the course of the three test periods. In addition, the users were learning how to use and adapting to the system. A primary goal, then, of longitudinal analysis like that performed in this case study is to learn how to monitor system performance (and what to monitor) in order to identify significant problems and direct efforts to the resolution of those issues. Our experience with this system suggests that it is critical to use subjective evaluations (e.g., user surveys) and objective criterion (e.g., ASR performance) to achieve this goal. This section describes three examples of the use of this combined information to guide system modifications.

For example, looking at both the ASR performance and the user surveys highlighted a discrepancy between the objective ASR performance and user perceptions. In the test scenarios, users were told to try to use the "cancel" command to negate a previous action. Twenty-seven percent of users reported that the system sometimes did not recognize the command and did not undo the previous action. However, as the ASR performance results in Table 2 show, recognition success for the "cancel at dialing" grammar on in-vocabulary words was very high (95%). In addition, there were very few rejected utterances for this grammar. This puzzling inconsistency between the data sources can be explained by the way the application and the recognizer communicated: when the recognizer detected energy that was deemed to be "garbage" it would terminate and return control to the application. As a result, the system was often not "listening" when the user said "Cancel" The fact that 34.9% of the input captured by the recognizer when the "cancel at dialing" grammar was active contained no speech probably reflects this problem, at least in part. However, without the subjective data, we would not have discovered the problem, because the default system action on receiving no speech input when the "cancel at dialing" grammar is active is simply to place the call. Under regular usage periods, we expect this outcome to be the "normal" pattern for a successful interaction.

Another problem that was diagnosed only after looking closely at multiple data sources was the cause of the relatively poor user perception of "recognition" accuracy for 10-digit account numbers during initial system access for both the scenario tests and regular usage. Our initial hypothesis was that the digit acoustic models were not adequate, but the labeled speech utterances identified three possible causes. First, the speech captured by the system often contained audible remnants of the not-completely canceled agent prompts, which the recognizers often rejected. Second, the initial time-outs of the system were set too short, so that in many cases the digit strings that the recognizer received were incomplete, and so not legitimate account numbers. Third, the beam-width parameter on the recognizer was quite narrow. All three of these factors resulted in a rejected utterance, leading to reprompts.

Finally, analysis of the out-of-vocabulary utterances demonstrated three major issues. Many of the OOV utterances were commands that were similar to available system commands but had not been included in the vocabulary (e.g., "goodbye" vs "goodbye Annie"), or items that were similar to phrases in the personalized grammars (e.g., "Candy Kamm" vs. "Candace Kamm"). We addressed the first problem by expanding some of the system grammars to include the most frequent OOV items. The second problem was alleviated by using a web page interface to voice label administration that allowed users to enter multiple aliases for a given phone number. Another significant proportion of the OOV items were commands spoken when the vocabulary that included those commands was not currently active - that is, a user orientation problem. To address this problem, we are exploring the use of auditory "earcons" to help orient the users to their "location" in the application feature space[4].

6. DISCUSSION

The ultimate goal for successful agent-based interactive systems is to define the optimal mapping between these user perceptions of success and system behavior, and to use this mapping to automatically adapt the system to increase user satisfaction. Our analysis indicates that the correlations between user satisfaction ratings and ASR performance measures are relatively weak, in contrast to our initial hypotheses. This brings up several fundamental issues related to probing users about their perceptions of and satisfaction with particular aspects of system performance. First, to what extent can users provide reliable judgments pertaining to their reactions to specific subcomponents of a system or aspects of an interaction with the system? Perhaps user ratings reflect a holistic view of system performance, and so users may not be able to provide judgments about any specific factor (e.g., ASR performance, system response time, or system reliability) independent of their overall experience with the system. As a result, user satisfaction measures might not be expected to correlate particularly well with any specific factor or component. Rather, user satisfaction might correlate better with measures that look at system performance more broadly - for example, relative frequency of usage of the system over time. Second, to what extent can users provide judgments that are localized in time to just the interactions that are captured in the analysis of the objective data? Obviously, user satisfaction judgments are based on the user's experiences with the system, integrated over time (with some decay). As a consequence, subjective measurements must be obtained from the user in a timely yet unintrusive manner. In this study, this goal was accomplished by capturing objective data for the period immediately prior to administering the survey, although the duration of the data collection period might not be optimal. Third, a number of our accounts had multiple users, but only one user per account completed the surveys. Users were not required to identify themselves, so it is difficult to automatically determine which utterances were spoken by the user who completed the survey. As a result, the system performance for an account is the average across users of that account, while the survey results will reflect only the perceptions of a single user. This situation is not unusual in a service that might be deployed on a per household basis, but it could have a significant impact on correlational analyses.

This case study represents our initial efforts to explore how objective measures of system performance can be related to subjective user perceptions to drive modifications of the system to yield better usability, task success, and user satisfaction. Although we have extended the evaluation data sources to include a wider set than typically studied, there are a number of other kinds of task- and system-related data that may be play a major role in influencing user satisfaction. In fact, in our proposed framework for system evaluation[3], we suggest that user satisfaction should be predictable as a function of a variety of objective cost factors and success measures, including but not limited to ASR performance, task success, system response time, number of system failures, and number of attempts required to task completion. Our future work will look at these and other factors in an attempt to better define the mapping between user perceptions of success and system behavior. Our current experiences demonstrate that achieving this goal is an iterative process. One of the main problems is that, in order to effectively make use of multiple data sources, considerable forethought and effort must be given to ensuring that the required objective and subjective data can be acquired, combined, and analyzed efficiently. Toward that end, we have proposed a unified database creation and management system for spoken dialog systems[5]. By continuing to probe subjective reactions to and objective performance of our system over time, we expect to refine our understanding of the principles that drive user perceptions and to apply them to the development of techniques for automatically improving human-machine interaction.

7. REFERENCES

[1] H. Meng, S. Busayapongchai, J. Glass, D. Goddeau, L. Hetherington, E. Hurley, C. Pao, J. Polifroni, S. Seneff, and V. Zue, "WHEELS: A conversational system in the automobile classifieds domain," in *Proc. of the Intl Conf. Spoken Lang. Processing*, vol. 1, (Philadelphia, PA), pp. 542–545, 1996.

[2] S. Bennacef, L. Devillers, S. Rosset, and L. Lamel, "Dialog in the RAILTEL telephone-based system," in *Proc. of the Intl Conf. Spoken Lang. Processing*, vol. 1, (Philadelphia, PA), pp. 550–553, 1996.

[3] M. Walker, D. Litman, C. Kamm, and A. Abella, "PARADISE: A framework for evaluating spoken dialogue agents," in *Proceedings of ACL/EACL97*, 1997.

[4] D. Dutton, C. Kamm, and S. Boyce, "Recall memory for earcons," in *Proc. of Eurospeech97*, (Rhodes, Greece), 1997.

[5] C. Lin, S. Narayanan, and R. Ritenour, "Database management and analysis for spoken dialog systems: Methodology and tools," in *Proc. of Eurospeech97*, (Rhodes, Greece), 1997.