VOICE-IF: A MIXED-INITIATIVE SPOKEN DIALOGUE SYSTEM FOR AT&T CONFERENCE SERVICES

M. Rahim, G.D Fabbrizio, C. Kamm, M. Walker, A. Pokrovsky, P. Ruscitti, E. Levin, S. Lee, A. Syrdal, K. Schlosser

AT&T Labs - Research
180 Park Avenue, Florham Park, NJ 07932.

Abstract

This paper presents the Voice-IF system; a mixed-initiative spoken dialogue system for AT&T conference services. One objective for creating Voice-IF is to provide a vehicle for evaluating our technologies in speech synthesis, recognition, understanding, dialogue and user interfaces on a real application with relatively novice users. Another objective is to design, build and test a set of tools that allow us to rapidly prototype spoken dialogue applications. In this paper, we describe the performance of Voice-IF during its 6-week deployment period. In particular, we report a) results of perceptual evaluations of the synthesized speech, b) system performance and user satisfaction ratings, c) PARADISE analysis of the data, and d) comparisons with other systems, including the W99 conference registration system used at the ASRU'99 workshop and the Travel Communicator system.

1. INTRODUCTION

There are tremendous opportunities for using natural language dialogue in automating a variety of limited-domain information-access based services, such as travel reservation and customer care. Building dialogue systems for such services requires extensive human resources from the initial data collection effort to the refinement steps of the application. Minimizing this effort by relying less on data and human expertise, thus shortening the development cycle for an application is clearly a challenge that needs to be addressed.

This paper describes the creation and development process of Voice-IF, a mixed-initiative spoken dialogue system that we deployed for the AT&T Innovation Forum. The goal of this effort is threefold: (1) to demonstrate portability and rapid prototyping of spoken dialogue systems using existing tools, (2) to evaluate our technology in speech synthesis, recognition, and spoken dialogue on relatively novice users, and (3) to provide a spoken language interface as an alternative to the web interface for users to register and access information about the Forum.

The AT&T Innovation Forum is a biannual event that provides AT&T employees and business customers a behind-the-scenes look at new technology and innovations from AT&T Labs. The Forum, which was held in November 2000, was two days long with the first day assigned for AT&T employees (Internal day) and the second day assigned for customers and sales executives (Customer day). Employees were able to register themselves or their customers for the Forum by using either Voice-IF or the Forum web site. Information regarding Voice-IF and a pointer to the web site were distributed to all employees through a company-wide mass e-mail.

In this paper we report on an evaluation of the spoken dialogue system based on interactions for 241 employees who used Voice-IF during the 6-week deployment period. Specifically, we present perceptual evaluations of the synthesized speech, evaluation metrics that describe the performance of the system and user satisfaction ratings, an analysis using the PARADISE framework [10], and comparisons of the results of Voice-IF with both the W99 conference registration system that was deployed in the ASRU'99 workshop [7] and the AT&T Travel Communicator system [3].

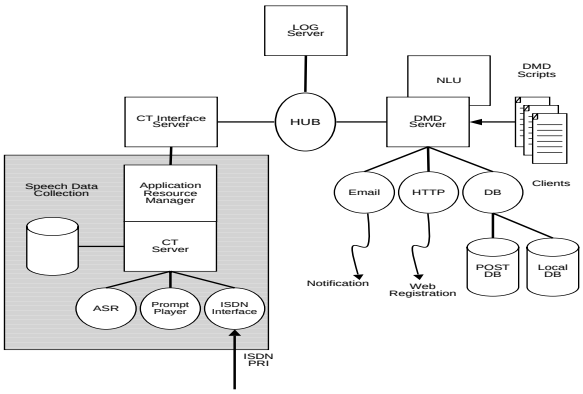

Figure 1: Architecture for the Voice-IF system.

2. SYSTEM ARCHITECTURE

The architecture for the Voice-IF system, shown in Figure 1, is based on the Hub-centric DARPA Communicator dialogue framework [8]. The Hub coordinates communication between the different servers, routes messages to a rule-based script, and logs the message traffic to a centralized location. The telephony sub-system (gray area in Figure 1) handles basic telephony capabilities like answering the phone call, sending audio to the speech recognizer, playing recorded prompts, detecting DTMF and facilitating speech data collection and logging. The Computer Telephony (CT) server [2] is an ECTF-compliant system that integrates the ASR and TTS components. The application Resource Manager (ARM), in combination with a command interpreter, translates Hub directives to internal function calls and synchronizes the thread of execution. All calls are routed through a network-grade echo canceler to remove electrical echo.

2.1. Automatic Speech Recognition (ASR)

The ASR system was based on the AT&T Watson continuous speech recognition technology [9]. The engine supports partial results and audio barge-in capabilities

both for stochastic as well as rule-based grammars. In the design of Voice-IF, the recognizer applied context-dependent hidden Markov models (HMM) with special phone set dedicated for digit recognition. The HMMs were trained on over 200 hours of speech using maximum likelihood estimation followed by minimum classification error training. Seven stochastic n-gram language models were incorporated, trained on a total of 5000 sentences. The ASR system maintained invariance to extraneous events such as clicks, pops, background noise and whistles by using dedicated garbage models. The system was also capable of producing confidence scores, which were compared to a predefined threshold for phrase acceptance/rejection.

2.2. Text-to-speech (TTS)

The system prompts were generated using the AT&T NextGen TTS [1]. The unit selection based system is trained on a female voice using an inventory of 6 hours of speech. To enhance the quality of the system, the database was greatly expanded to include an additional 6.85 hours of speech composed primarily of read scripts representing actual interactive dialogues between human agents and customers. Both the general-purpose and the expanded systems were capable of synthesizing any given English utterance.

2.3. Spoken Language Understanding (SLU)

The AT&T CHRONUS SLU system was adopted as a Hub server [4]. The system receives data structures containing sentence hypotheses and returns templates that include semantic interpretations. No complete syntactic analysis is carried out, partially due to the lack of in-domain training data.

2.4. Dialogue Manager (DM)

The dialogue manager was based on the AT&T DMD [5] system. The DM applies a scripting language that represents the current state of the dialogue and generates new templates that include the request for the next dialogue action. Dialogue actions contain either text strings for TTS, grammar pointers for ASR or requests for database inquiries.

3. DIALOGUE STRATEGY AND APPLICATION

During the initial four weeks of the Voice-IF deployment, the system greeted callers with the prompt "*Hi! Welcome to the Innovation Forum. You may register or get information about the Forum. You may say help at any time. What can I do for you?*" Close to the completion of the service, the greeting prompt was changed to the open-ended prompt "*Hi! Welcome to the Innovation Forum. How May I Help You?*". This modification allowed us to study the effect of the specificity of the opening prompt on task completion and user satisfaction.

The functionalities supported by Voice-IF included (1) registration for the Internal day, (2) registration for the Customer day, (3) registration for both days, (4) general information, (5) information about the Internal day, (6) information about the Customer day, (7) technical agenda, (8) social events, (9) hotel information, (10) directions, (11) local restaurants, (12) dates, and (13) general help (including a pointer to the Forum web site). All these functionalities were available to callers at the top menu in the dialogue strategy.

3.1. Registration

The Innovation Forum web registration form requires personal information that includes name, email address and affiliation. Attendees are also requested to provide their preferred day of attendance, whether or not they are attending in person or via video-conference, whether or not they are bringing a customer, and their lunch preference. To retrieve personal information, Voice-IF asks employees for their Human Resource ID number. The DM has been extended to execute SQL queries that access a relational database of AT&T corporate personnel information that includes over 140,000 employee profiles. A second local database is used to store information on registered attendees.

Registration using Voice-IF mimics the traditional web registration process. Once the conference registration data is collected, it is submitted to the Innovation Forum web site. This feature allows the service to interoperate with the existing web interface without interfering with the actual registration process. The HTTP client in Figure 1 is responsible for filling out the web form and posting it to the web site. The operation is completely transparent to the web manager, which is not able to distinguish between registration that is completed over the phone or through the web. A notification to the user summarizing the result of the transaction is sent by the web manager.

3.2. Information Access

In addition to registration, Voice-IF can provide callers with information about both the Internal and Customer days, with an option to receive more extensive information and register directly for either event. The system can also provide information about hotels, directions and local restaurants. More extensive information can be provided to the caller via email upon request. In this situation, the caller's email address is retrieved by launching a database query using the callers' ID number - a similar process to that used during registration.

3.3. Farewell

Before ending the call, users were asked the question *Did you like using the system?* The response to this prompt was used as a measure of user satisfaction.

Voice-IF was designed so that callers could generally take initiative after an explicit confirmation by the system. However, when the user's input was either silence or contains invalid information, Voice-IF changed to a system initiative mode with more specific prompts and constrained language models. Prompts, generated using the dialogue-domain expanded TTS, were customized for the Voice-IF system to target infrequent novice users and to present a friendly, informal agent personality.

4. PORTABILITY

Voice-IF took two months to design, build, test and deploy. The telephony platform, technology components and tools to build the application were identical to those used in the Travel Communicator system. Similar tools

were also used in building the W99 system that was deployed for the IEEE sponsored ASRU'99 workshop.

The creation of W99 and the Travel Communicator systems were instrumental for fast prototyping Voice-IF. Due to the lack of in-domain data, the acoustic models were trained on speech from the Travel Communicator, W99, and other in-house customer care applications. The language models were primarily based on those of W99 with appropriate modifications made to accommodate for new vocabulary phrases. Unlike the W99 system, however, both the SLU and the DM were largely extended to accommodate for the more complex registration and information access mechanisms implemented in Voice-IF. These extensions, and the application-specific prompt generation process consumed the majority of the Voice-IF development effort.

5. EXPERIMENTAL EVALUATION

In this section, we present four sets of evaluations; (1) perceptual evaluation comparing the general-purpose TTS versus the dialogue-domain expanded TTS, (2) system performance over the 6-week deployment period, (3) a PARADISE analysis and model of the system, and (4) comparative evaluation of the Voice-IF system with the W99 and the Travel Communicator systems.

5.1. TTS Evaluation

Two types of perceptual evaluations comparing the baseline TTS with the dialogue-domain expanded TTS demonstrated statistically significant differences. In a blind paired comparison listening test, 7 listeners heard 11 pairs of Innovation Forum user-system turns; a version of each sentence was synthesized by each of the two TTS systems. Listeners preferred sentences synthesized by the dialogue-domain TTS over the baseline TTS 74% of the time. A formal subjective listening test was conducted in which 40 naive listeners independently rated the overall quality of 10 test sentences synthesized by each system using a 5-point scale. The test sentences synthesized by the dialogue-domain expanded TTS were rated 0.27 points higher than the same sentences generated by the baseline TTS.

5.2. Voice-IF Evaluation

241 calls were received during the 6-week deployment phase of the system. Call logs were post-processed to identify which subtasks had been accomplished and to compute objective metrics including User Turns, Task Duration, System Words/Turn, User Words/Turn, and Word Accuracy on a per dialogue basis.

User satisfaction was assigned on a 4 point scale based on hand transcription of the survey responses. Positive responses, e.g. *Yes, it was great*, were assigned a value of 3. Nonenthusiastic positive responses, e.g. *It was okay*, were assigned a value of 2. Negative responses were assigned a value of 1. The value 0 was reserved for those callers who terminated the call before answering the survey question.[1]

The mean results for the registration subtasks are shown in Table 1. Not unexpectedly subjects who completed a registration task (Both, Internal and Customer) had more User Turns and longer Task Durations than subjects who did not register. This may reflect the fact that users who did not register (Neither) had more difficulty with the system as evidenced by lower Word Accuracy, and lower user satisfaction. Note that there don't appear to be significant differences in any measures for users registering for just the internal day or just the customer day. However, interestingly, users who registered for *both* days had lower Word Accuracy and user satisfaction than users who only attempted one registration subtask. A plausible explanation for this difference was that the dialogue flow for navigation between the registration subtasks was non-intuitive for users, resulting in some user confusion and subsequent lower satisfaction. It should be pointed out that the false acceptance rate for registration was zero.

Metric	Both Days (N=49)	Internal Day Only (N=121)	Customer Day Only (N=21)	Neither Day (N=50)
UserTurns	13.6	12.0	13.2	6.9
TaskDur/sec	201.5	170.3	192.3	98.5
SystemWords/Turn	19.6	19.7	19.7	22.3
User Words/Turn	3.0	2.7	2.5	4.7
Word Accuracy	79.8	83.6	85.0	73.3
UserSat	1.9	2.4	2.3	0.4

Table 1: Means per Registration Subtask for 241 Calls

5.3. PARADISE Evaluation

Factor	Coefficient
Registration Task	0.34
UserTurns	0.24
Word Accuracy	0.16
SystemWordsPerTurn	-0.12
Customer Day	-0.14

Table 2: PARADISE Results

We developed models of user satisfaction as a function of other metrics by applying multivariate linear regression as per the PARADISE framework [10]. Factors that are significant predictors of user satisfaction and the magnitude of their contribution is provided in Table 2. The five significant factors were completion of the registration task (either Internal or Customer), completion of Customer day task, User Turns, System Words/Turn and Word Accuracy, accounting for 35.6% of the variance in user satisfaction. Our interpretation of these factors follows. Longer interactions (more user turns) were indicative of successful users, because other users hung up on the system. Successful registration is an indication of happier users, but this is offset slightly by the navigation difficulties of those users who registered for both days (accounting for the negative coefficient for Customer Day). System Words Per Turn, which negatively affected user satisfaction, can be viewed as an indirect indication of ASR rejections, because rejection utterances on average were much longer than other utterances; typically they included either an apology or additional in-

[1] Our assumption is that the callers who terminated early had trouble with the system. This assumption is strongly supported by objective metrics demonstrating that ASR performance was much worse for these calls.

structions. Finally, as usual, higher Word Accuracy led to greater user satisfaction.

Metric	DIRECT (N=206)	OPEN (N=35)
User Turns	11.4	11.4
Word Accuracy	81.3	77.8
User Words Turn	3.0	4.2
user satisfaction	1.9	1.6

Table 3: Comparison of Direct and Open-ended Prompts

As mentioned above we modified the initial prompt from a directive prompt to an open prompt for the last 15% of the data collected. The effect of this prompt change was to increase the average number of User Words per Turn, resulting in a decrease in both Word Accuracy and user satisfaction, as summarized in Table 3. This result is consistent with other evaluation results [6].

5.4. Comparative Results

Metric	W99 (N=550)	Voice-IF (N=241)	Communicator (N=81)
UserTurns	4.8	11.4	20.5
Word Accuracy	68.0	80.8	72.73
UserWordsPerTurn	4.6	3.8	2.3

Table 4: Comparison with the W99 system

Finally, it is interesting to compare the Voice-IF system with the W99 conference registration system and the Travel Communicator System. As shown in Table 4, the Voice-IF interactions were on average twice as long as those of W99, with fewer User Words Per Turn and higher Word Accuracy. The longer interactions were attributed to the more complex registration process in Voice-IF as compared to that of W99. The higher word accuracy can be attributed to the lower user words per turn and the more advanced acoustic and language models used. Compared to the Communicator system, Voice-IF has on average half the number of interactions, higher word accuracy and higher user words per turn. The longer interactions in Communicator are attributed to the complex reservation process that included both air and ground reservations. The lower word accuracy in Communicator is due to the difficulty in recognizing city names, dates and times which constituted the majority of the data collection.

6. SUMMARY

One interesting area in spoken dialogue research is the problem of rapid prototyping of complex mixed-initiative systems. The challenge here is not only to create the technology with general-purpose tools that would facilitate portability and reusability of data and modules across different domains, but also to provide the infrastructure that would enable integration of models and components in a robust and a rapid manner. As part of this effort of rapid prototyping, we presented in this paper the development of the Voice-IF system that was deployed for the AT&T Innovation Forum. The system was used as a vehicle to evaluate our speech and natural language dialogue technology.

This paper presented the results of the 6-week deployment which demonstrated the following. (1) On a 5-point scale, the dialogue-domain expanded TTS was rated 0.27 points higher over the general-purpose AT&T female voice. (2) On a 4-point scale (where 3 was considered excellent), user satisfaction for registered users was rated between 1.9-2.4, with corresponding word accuracy between 79.8%-85%. However, users who did not register had more difficulty with the system as evidenced by lower word accuracy of 73.3%, and lower user satisfaction rating. (3) Consistent with previous studies, PARADISE analysis demonstrated that user satisfaction with the system was highly influenced by speech recognition performance and successful completion of the registration task. (4) Interestingly enough, we found that using an open-ended prompt, as opposed to a directed prompt, resulted in lower user satisfaction. (5) Voice-IF had a higher word accuracy and twice the number of interactions than the W99 system. Finally, we should point out that Voice-IF registered over 28% of attendees of the Internal day.

Acknowledgments

The authors would like to acknowledge the technical contribution of Dawn Dutton, Shrikanth Narayanan, Giuseppe Riccardi and Ralph Knag. This project was partially funded by DARPA under the Communicator project number MDA972-99-3-0003.

7. References

[1] M. Beutnagel, A. Conkie, J. Schroeter, Y. Stylianou, and A. Syrdal. The AT&T next-gen tts system. In *Joint Meeting of ASA, EAA and DAGA*, 1999.

[2] G. Di Fabbrizio, C. Kamm, P. Ruscitti, S. Narayanan, B. Buntschuh, A. Abella, J. Hubbell, and J. Write. Extending a standard-based ip and computer telephony platform to support multi-modal services. In *Workshop on Interactive Dialogue in Multi-modal Systems*, pages 22–25, 1999.

[3] E. Levin, , S. Narayanan, R. Pieraccini, K. Biatov, E. Bocchieri, G. Fabbrizio, W. Eckert, S. Lee, A. Pokrovsky, M. Rahim, P. Ruscitti, and M. Walker. The AT&T darpa communicator mixed-initiative spoken dialogue system. In *ICSLP*, 2000.

[4] E. Levin and R. Pieraccini. CHRONUS, next generation. In *Proc. ARPA Spoken Language System Workshop*, January 1995.

[5] E. Levin, R. Pieraccini, W. Eckert, P. Di Fabbrizio, and S. Narayanan. Spoken language dialogue: From theory to practice. *Submitted to IEEE ASRU Workshop*, December 1999.

[6] S. Narayanan, G. DiFabbrizio, J. Hubbell, C. Kamm, B. Buntschuh, Ruscitti P., and J Wright. Effects of dialog initiative and multimodal presentation strategies on large directory information access. In *ICSLP*, pages 16–20, 2000.

[7] M. Rahim, R. Pieraccini, W. Eckert, E. Levin, G. DiFabbrizio, G. Riccardi, C. Kamm, and S. Narayanan. A spoken dialog system for conference/workshop services. In *ICSLP*, pages 16–20, 2000.

[8] S. Seneff, R Lau, C. Pao, P. Schmid, and V. Zue. Galaxy ii: A reference architecture for conversational system development. In *ICSLP*, pages 931–934, 1998.

[9] R.D. Sharp, E. Bocchieri, C. Castillo, S. Parthasarathy, C. Rath, M. Riley, and J. Rowland. The Watson speech recognition engine. In *Proc. Int. Conf. Acoust., Speech, Signal Processing*, pages 4065–4068, 1997.

[10] M. Walker, C. Kamm, and D. Litman. Towards developing general models of usability with PARADISE. *Natural Language Engineering: Special Issue on Best Practice in Spoken Dialogue Systems*, 2000.