A SPOKEN DIALOGUE SYSTEM FOR CONFERENCE/WORKSHOP SERVICES

Mazin Rahim, Roberto Pieaccini, Wieland Eckert, Esther Levin,
Giuseppe Di Fabbrizio, Giuseppe Riccardi, Candy Kamm, Shrikanth Narayanan
AT&T Labs-research, 180 Park Avenue, Florham Park, NJ.

ABSTRACT

This paper describes our progress towards building a telephony-based spoken dialogue system for workshop/conference services. A mixed-initiative dialogue system has been developed that is engineered to offer users natural interaction with the system, ease-of-use and robustness towards ambiguous requests and machine errors. A prototype system, known as W99, is described in this paper which was deployed in the 1999 IEEE International Workshop on Automatic Speech Recognition and Understanding (ASRU'99), Keystone, Colorado. This system integrates advanced technologies in speech and dialogue design. An evaluation of the W99 system in terms of recognition performance, understanding accuracy and dialogue success rate during the live trial of the system are presented in this paper.

1. INTRODUCTION

Advances in computing power and speech and language processing technologies have opened tremendous new opportunities for using spoken dialogue systems in real-world applications. Several of these applications have evolved in recent years that use natural language dialogue for automating a variety of complex services such as train information [4], travel reservation [5], and customer care [3]. This paper reports on our progress towards building a telephony-based spoken dialogue system for automated workshop/conference services. The objective is to design a system that would guide both *novice* and *expert* users through conference services *robustly* and *intelligently*, producing reasonable responses, even when the user's query is not within the scope of the application. The main functionalities of the designed system are automated conference services and information access. This includes conference registration, hotel information and reservation, and various other services such as information about the paper status, paper submission, technical program, social events, location, dates, web site, transportation and fees.

In this paper we describe the development of the W99 system - a mixed-initiative dialogue system for conference services that was deployed in the ASRU'99 workshop. In [8], we presented the performance of W99 on a controlled experiment with 50 subjects exercising four different scenarios. This paper will address the main components and functionalities of the system and present the performance during the live trial from September 1999 and for a period of four months. An evaluation of the system in terms of recognition accuracy, concept accuracy and dialogue success rate are reported for 550 dialogue interactions.

2. SYSTEM ARCHITECTURE

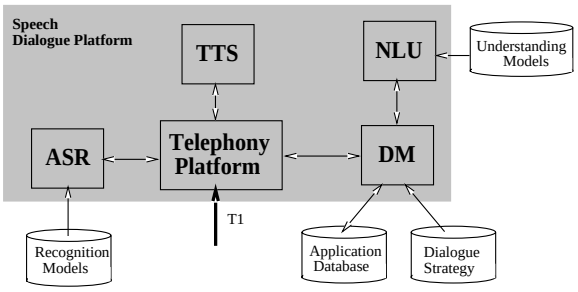

Figure 1. A simplified architecture of the W99 system.

A simplified architecture of the W99 system is shown in Figure 1. The four major components of the system, namely, the ASR (automatic speech recognition) engine, DM (dialogue manager), TTS (text-to-speech) synthesizer, and the NLU (natural language understanding) module, are all interfaced through the Telephony Platform [2]. This is a standard open-platform dialogic hardware that connects to a T1 line. Although the four components are designed to be application independent, they use dedicated set of models for recognition, understanding, dialogue strategy and application database.

The ASR includes the AT&T Watson engine which is capable of providing both complete and incomplete recognition hypotheses in real-time [11]. The parameters of this engine (e.g., grammar identifiers) can be set dynamically while the system is in a "listening" mode.

The AT&T TTS system is based on unit selection [1]. It accepts text strings including prosodic markers and returns synthesized speech. This system provides natural and intelligible speech and was highly rated in the November 1998 ESCA/COCOSDA TTS comparison.

The NLU module is based on key-phrases being associated with relevant concepts. It receives data structures (or templates) containing sentence hypotheses and returns templates that include their semantic interpretation [6]. No complete syntactic analysis is carried out partially due to the lack of sufficient training data.

The dialogue manager is implemented as a DMD (Dialogue Manager Development) script. This scripting lan-

guage was developed at AT&T for the implementation of dialogue strategies [7]. The DM processes templates that represent the current state of the dialogue and generates new templates that include the request for the next dialogue action. Dialogue actions include text strings for the TTS, grammar pointers for the ASR or requests for database queries.

For the remainder of the paper, we will describe the dialogue strategy and the ASR development in more details.

3. DIALOGUE SYSTEM

3.1. Functionalities

The main functionalities for conference services are registration, technical submission and information access. Registration through W99 is limited to members of the IEEE Signal Processing Society (SPS).[1] As the majority of the participants are IEEE members then accessing their profile by automatically recognizing their membership number, as opposed to their names and addresses, is both easy and accurate. Non-IEEE members are directed to register through the web. Whether users register through W99 or the Web, they receive an email containing their private four-digit code. This code is essential for accessing a variety of services, including checking paper status and changing user profile.

In addition to registration and checking paper status, W99 also provides basic information concerning hotels and fees, registration costs, transportation, dates and times, technical agenda and social events.

3.2. Dialogue Strategy

The initial stage in building a system for conference services included developing a web-based prototype that uses text input. The dialogue strategy and the functionalities of the system were then tailored upon these responses which were also applied for building language models for ASR.

The W99 system adopts a mixed-initiative dialogue strategy that is engineered to provide three essential features:

Naturalness: The ability for users to converse with the system in an open dialogue environment is essential in providing natural human-machine dialogue. In W99, users have the flexibility to speak fluently to the system on issues related (or not) to conference services. Key-phrases are identified in the form of attribute-value pairs from users' input and the most appropriate dialogue strategy is executed [7]. In case of an unreasonable request, W99 directs the user to the workshop web-site or provides a telephone number for further information.

Besides adopting an open dialogue structure, another important feature for natural language dialogue is allowing callers to interrupt the system at any time while the prompt is playing. This is referred to as *barge-in*. In W99, barge-in is enabled at the appropriate key-phrases that are associated with semantic concepts. During false barge-in, i.e, system interruption but with an invalid response, W99 switches to a system-initiative mode.

[1] A copy of the IEEE SPS database which includes over 22,000 members was provided to us courtesy of Mercy Kowalczyk.

W99 uses prompts that are automatically generated from TTS. The quality of the synthesized speech plays an important role in simulating a natural dialogue interaction. The high quality synthesis, the mixed-initiative dialogue strategy, the low latency (time-to-first audio) and the barge-in capability all together contribute significantly to improved naturalness when conversing with W99.

Ease-of-use: We define ease-of-use as the ability for both novice and expert users to access information in a straightforward manner. The functionalities in the W99 system are designed in a tree form where the root node corresponds to the greeting prompt. The number of branches corresponds to the different functionalities of the system. Quick and easy access to these functionalities at any point in the interaction are supported through context switching. To help navigate through the tree, especially in periods of system error, various levels of guidance are provided in the form of help prompts.

Other factors that play a role in improving the intelligibility of the dialogue is the information contents of the system responses. Besides offering short and informative prompts, W99 can accommodate for ambiguous and recurrent requests.

Robustness: The ability to maintain natural and constructive dialogue is clearly a challenge when dealing with spoken dialogue systems as opposed to text-based dialogue systems. Without sufficiently large training corpus, the diverse and unpredictable set of responses and background environments from both novice and expert users would pose a robustness problem at all levels of the system including acoustic, language, understanding and dialogue. Robustness at the acoustic and language levels will be discussed in the next section.

To achieve some level of robustness at the dialogue level, W99 is designed to switch from a user-initiative mode to a system-initiative mode in the event of user or machine error. This is typically identified when a key-phrase is either missing or has a low confidence score. For example, to activate the concept PAPER_STATUS the system needs to detect the phrases paper and status. Should status be misrecognized as "six us", for example, the dialogue would recover as follows:

Recog: Need to know six us paper
W99: Would you like to know about the call for papers?
Recog: Not really
W99: Would you like to know about the status of your paper?
Recog: You bet
W99: OK. I can help you with that. Do you have the access ....

4. ASR SYSTEM

4.1. Acoustic Modeling

Due to the lack of in-domain data, early deployment of the W99 system in July 1999 (Phase 0) included an off-the-shelf acoustic models from the *How May I Help You* (HMIHY) study [9]. These models consisted of two sets of sub-word units; one dedicated for the digits and the other for the remaining vocabulary words. Each set applied left-to-right continuous-density hidden Markov models (HMMs) with unit durations that were approximated by a gamma

distribution. These HMMs have been trained using maximum likelihood estimation (MLE) followed by minimum classification error (MCE) training [9].

The W99 system was updated in August'99 (Phase 1) and September'99 (Phase 2). In each deployment, acoustic data from the previous phase was used to further enhance the HMMs through MCE training. For each of the Phase 0, Phase 1 and Phase 2, there were 750, 2095 and 3550 sentences that were collected, respectively.

An important element in the development of robust spoken dialogue systems is maintaining invariance to extraneous events, such as clicks, pops, background noise, echos, whistles, etc. This is particularly important in W99 due to the barge-in capability which in some instances may cause extraneous events to be misrecognized as key-phrases, resulting in frequent system interruption and poor dialogue interaction. Besides garbage modeling, W99 performs online rejection in which a confidence score based on a likelihood ratio distance is computed and compared against a predefined threshold for phrase acceptance/rejection. The system is also equipped with a voice activity detector and a hardware acoustic echo canceler.

4.2. Language Modeling

Early incarnation of the W99 system applied a stochastic word bigram language model, borrowed from the HMIHY field-trial – a rather different application to W99 [3]. As data from the web-based dialogue system became available, they were used for incremental adaptation of the language model. Although the majority of the data did not truly capture the spontaneous nature of speech input, they represented an excellent seed for building language models.

Four distinct language models were employed in Phase 0, Phase 1 and Phase 2. These models were applied for "greeting", "confirmation", "digits" and "help". Each model was trained from a separate corpus of text data by using n-gram stochastic finite state automata[10]. With the exception of the "digits" model that was trained on the IEEE SPS membership directory and the access code database, the remaining language models were generated using up to 3100 sentences and a lexicon of 2300 words. Compound language models were also generated to accommodate for embedded digits in the dialogue.

5. SYSTEM EVALUATION

Two sets of experiments have been performed for evaluating the W99 system. The first experiment was conducted with 50 subjects, each was being asked to perform four scenarios that included registration and information access [8]. The second experiment, involved 550 dialogue interactions that were recorded during the live trial.

5.1. Experiment I

Details of this experiment are presented in [8]. Subjects were asked to perform four different scenarios which included workshop registration, information on hotel fees and directions, information on paper submission and finally one open scenario of the subject's choice.

Two sets of questionares were completed by each subject. The first included a measure of the "task success rate" for each of the four functionalities. This resulted in 60-96%. The second set of questions were a subjective evaluation of users' interaction with W99. The intent was to evaluate the system's performance specifically in the areas of naturalness, ease-of-use and robustness [8].

Scoring was tabulated from 1-5 with 1 being *very difficult,* or *very bad* and 5 being *very easy* or *very good.* Our results show subjects were generally satisfied that the system "understood" them, giving it an average score of 3.5 across all tasks. The lowest score, with an average of 3.0, was related to the ease-of-use of the system. An interesting result was an average of 3.5 being given to measure the robustness of the system and is ability to guide users through the application in a sensible manner. Finally, in terms of overall rating, W99 scored an average of 3.2, with the lowest score being assigned to the open scenario.

The performance of the system for Phase 0 and Phase 1 are illustrated in Table 1. Performance was quantified in terms of word accuracy after insertion, deletion and substitution errors, and concept accuracy which is the discrepancy in the NLU output when using the recognized speech as opposed to the transcription. As pointed out earlier, both acoustic and language models were incrementally adapted during each phase of system deployment. The results show that although the word accuracy is 53.3% for Phase 1, which is only slightly better than that for Phase 0, the concept accuracy is at 74.8% with the out-of-vocabulary rate being at 1%. It is interesting to also note that for this operating point, users' overall rating of W99 was 3.2.

	WA	CA
Phase 0 (Jul'99)	50.2	70.4
Phase 1 (Aug'99)	53.3	74.8
Phase 2 (Sep-Dec'99)	68.0	80.0

Table 1. The performance of the W99 system in terms of word accuracy (WA) and concept accuracy (CA).

5.2. Experiment II

This experiment represents the live trial which involved callers using the W99 system for workshop registration, checking paper status and general information access. 550 actual dialogues were collected during the system's Phase 2 deployment in September, 1999. This reflected a total of 3550 user responses, 4% of which were international calls.

Since we are interested in the dialogue success rate as well as user satisfaction, there are typically several factors that may play a role here. Such factors include the recognition and semantic accuracy, average number of turns per dialogue, average number of words per dialogue turn, duration of each interaction, false rejection/acceptance for registration/paper status, percentage of calls with an ending note, percent of obtaining first/second help and percent of being disconnected by the system. Recent studies have shown that these factors can be integrated using machine learning techniques to successfully predict problematic situations within a dialogue [12].

Figure 2 provides histograms for the number of user turns per dialogue interaction and the number of words per dialogue turn. The average values for these statistics are 4.77 and 4.64, respectively. This suggests that the interac-

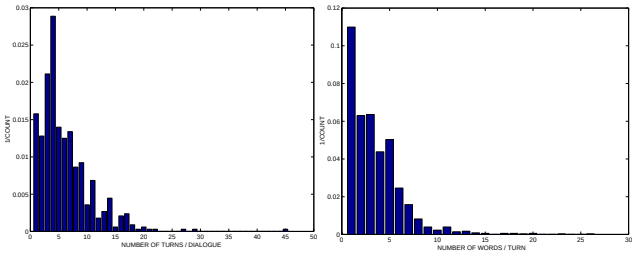

Figure 2. (a) Number of user turns per dialogue; (b) Number of words per dialogue turn.

tions are fairly short and equivalent to the AT&T HMIHY task[3].

Table 1 presents the word and concept accuracies during Phase 2 deployment which correspond to 68% and 80%, respectively. The improved results in Phase 2 may be attributed to the use of more accurate in-domain acoustic and language models from the Phase 1 trial. The high accuracy resulted in more successful interactions and lesser false acceptance rate which is especially critical for registration and checking paper status. Of the 60% attendees which were IEEE members, 14% registered through W99 with no false acceptance. Also of the 100 papers submitted, 60% were checked through W99 with no false acceptance.

Since the majority of users were first time callers, 65% of them either requested for help or were given help when their request was misunderstood by the system. 19% of calls included a second help prompt and 7% were ended by W99[2]. 44% of the dialogues finished with an ending note.

6. SUMMARY

This paper presented a spoken dialogue system for workshop/conference services. A prototype system, referred to as W99, was deployed for the IEEE ASRU'99 workshop for registration and information access. This system represents an important milestone at using advanced speech processing technologies for workshop/conference services, ranging from speech recognition and synthesis to dialogue design.

The W99 system is developed using a mixed-initiative open-dialogue structure, offering users natural interaction with the system, ease-of-use and robustness to ambiguous requests and recognition errors. The system provides high-quality TTS, fast response, barge-in capability and flexible NLU, which collectively contribute to a natural human-machine dialogue. In addition, W99 is equipped with discriminatively-trained acoustic models, a voice-activity detector, echo canceler, rejection and garbage modeling capabilities which all play a major role in maintaining robustness to extraneous events and changing environmental conditions.

Two experiments were reported in this paper. The first was a controlled experiment involving 50 subjects that were asked to perform four different scenarios. Scoring from 1-5, W99 achieved an overall subjective rating of 3.2, with a word accuracy of 53.3% and a concept accuracy of 74.8%. It also achieved an average score of 3.4 when callers where asked whether the system can be used as a complementary modality to web access for workshop/conference services.

[2]W99 ends any call after three consecutive help prompts

In the second experiment, 550 dialogue interactions were collected during the live trial. The system successfully automated 60% of paper status and 14% of registration with no false acceptance. The word and concept accuracies were 68% and 80%, respectively.

Given the open dialogue structure of this task, the limited data collection that we had and the limited time-frame in developing and deploying this application, we believe that the performance of the system and its ratings are an indication of a successful application using spoken language dialogue.

Acknowledgments

The authors thank V. Goffin, R. Knag and M. Beutnagel for their technical contribution. This study was partially funded by the DARPA Communicator project MDA972-99-0003.

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] A.L. Gorin, G. Riccardi, and J.H. Wright. How May I Help You? *Speech Communication*, 23:113–127, 1997.

[4] L. Lamel, S. Rosset, J.-L. Gauvain, and S. Bennacef. The LIMSI ARISE system for train travel information. In *Proc. Int. Conf. Acoust., Speech, Signal Processing*, 1999.

[5] E. Levin, , N. Narayanan, and R. Pieraccini et. al. The AT&T darpa communicator mixed-initiative spoken dialogue system. In *ICSLP*, 2000.

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

[7] 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.

[8] M. Rahim, R. Pieraccini, W. Eckert, E. Levin, G. Di Fabbrizio, G. Riccardi, C-M. Lin, and C. Kamm. W99 – a spoken dialogue system for the asru'99 workshop. In *Proc. IEEE ASR Workshop*, December 1999.

[9] M. Rahim, G. Riccardi, J. Wright, B. Buntschuh, and A. Gorin. Robust automatic speech recognition in a natural spoken dialogue. In *Workshop on Robust Methods for Speech Recognition in Adverse Condition*, Tampere, Finland, 1999.

[10] G. Riccardi and A. GOrin. Stochastic language adaptation over time and state in natural spoken dialog systems. *IEEE Transactions on Speech and Audio Processing*, 8(1):3–10, January 2000.

[11] 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.

[12] M. Walker, I. Langkilde, J. Wright, A. Gorin, and D. Litman. Learning to predict problematic situations in a spoken dialogue system: Experiment with how may i help you? In *Proc. North American meeting of the association for computational linguistics*, 2000.