W99 – A SPOKEN DIALOGUE SYSTEM FOR THE ASRU'99 WORKSHOP

Mazin Rahim, Roberto Pieraccini, Wieland Eckert
Esther Levin, Giuseppe Di Fabbrizio, Giuseppe Riccardi, Chih-mei Elaine Lin, and Candy Kamm

AT&T Labs - Research
180 Park Avenue, Florham Park, NJ 07932.
{*mazin,roberto,eckert*}*@research.att.com*

ABSTRACT

This paper describes the development of W99 – a spoken dialogue system that is used in the Automatic Speech Recognition and Understanding (ASRU'99) workshop for registration, checking paper status and limited information access. W99 adopts a mixed initiative open dialogue structure, offering users natural interaction, ease-of-use and robustness. The system integrates advanced technologies in speech synthesis and recognition, dialogue design and user-interface. An evaluation of the W99 system in terms of recognition performance, understanding accuracy and dialogue success rate is presented in this paper.

1. INTRODUCTION

Advances in computing power and speech processing technologies have opened tremendous new opportunities for using voice-enabled systems in real-world applications. In this study, we report on our progress towards developing a telephony-based spoken dialogue system for general workshop/conference services. A prototype system, referred to as W99, has been developed and is currently being used at the ASRU'99 workshop for registration and limited information access.[1]

An important criterion in the design of the W99 system is the ability to converse with users in a *natural* open-dialogue environment on issues related to workshop services. This presents several new challenges to dialogue design and user-interface, especially as the majority of users have no familiarity or prior training of the system. The success of W99 will also be determined by its *ease-of-use*. This is particularly important since participants can, in principle, obtain similar information as that provided by W99 and perhaps more through web access.

Another important requirement in the design of the W99 system is *robustness*. At the acoustic level, this implies that variations in the acoustic characteristics of the speech signal due to extraneous conditions (such as different microphone handsets or background noise) should have little or no degradation on the performance of the recognizer. At the language level, users should be able to express themselves naturally and freely without being trapped by the constraints imposed by the language model. At the understanding level, the presence of disfluencies (such as ah, mm, etc) and recognition errors should have no or little impact on the behavior of the system. Finally, robustness at the dialogue level implies that the dialogue manager should guide users with different levels of expertise through the application seamlessly and intelligently. Maintaining robustness at these various levels is the key to the success of spoken dialogue systems in general.

In this paper, we describe the major components and the development process of the W99 system. We address the challenges in building an application for workshop/conference services without prior data collection or task-specific acoustic and dialogue models. The performance of W99 in terms of recognition and dialogue success rates will be reported on an evaluation set in which 50 subjects, whom many had little or no technical background in speech processing, were asked to converse naturally with W99.

2. W99 SYSTEM OVERVIEW

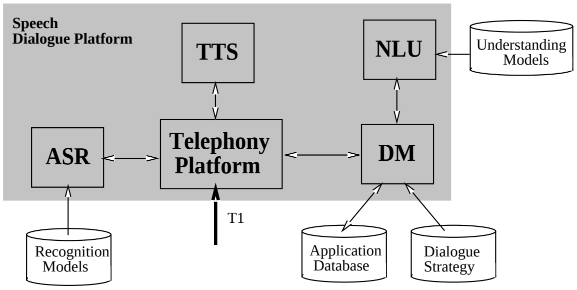

Figure 1. A simplified architecture of the W99 system.

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 [6]. 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 highly 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. No

---

[1] Visit http://asru99.research.att.com for further information.

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 language was developed at AT&T for the implementation of dialogue strategies [4]. 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 ARCHITECTURE

3.1. Functionalities

The main functionalities of the W99 system are registration, checking paper status and general information access.

Registration through W99 has been limited to IEEE Signal Processing Society (SPS) members.[2] Since the majority of participants are typically IEEE members then accessing their profile by automatically recognizing their membership number is both easy and accurate. Users who register through W99 receive an email containing their private access code. This code is essential for all communication, including checking paper status, changing user profile, etc.

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

3.2. Dialogue Strategy

The initial stage in the design of the W99 system included building a web-based prototype that uses text input. The dialogue strategy and the functionalities of the system were tailored based on users' responses. These responses were later used for building language models for ASR.

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

Naturalness: This is a critical feature in developing spoken dialogue systems. It affects largely people's perception and their overall satisfaction of these systems, and accordingly, their willingness to use them.

The most essential component in providing natural human-machine dialogue is the provision of an open dialogue structure. In W99, users have the flexibility to speak fluently and naturally to the system on anything relevant (or irrelevant) about the workshop. Key-phrases are identified from users' requests and the most "reasonable" dialogue strategy is executed. In case of an unreasonable or foreign request, W99 directs the user to the workshop website or provides a telephone number for further information. W99 is also capable of disambiguation as illustrated in the following:

W99: Welcome to the ASRU'99 voice automated system. What can I do for you?
User: Hi, I'm calling to find out about general costs
W99: Would you like to know about the cost of registration?
User: No
W99: Would you like to know about the cost of the hotel?
User: Sure that will be fine
W99: The cost of the hotel room ranges ...

Besides adopting an open dialogue structure, another important requirement for natural spoken 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 system-initiative with an explicit confirmation. For example,

User: Yep, I need to register
W99: I can help you with registration. Are [BARGE-IN]
User: Huh
W99: Are you an IEEE signal processing society member, please say either yes or no.

W99 uses prompts that are automatically generated from TTS. The quality of the synthesized speech plays an important role in simulating a human-to-human-like natural dialogue interaction.

Ease-of-use: We define ease-of-use as the ability for users to access information in a straightforward manner by being guided seamlessly and intelligently through the application. This is particularly important in W99 since most callers may use the system once or twice without prior training or knowledge of its functionalities.

In W99, the dialogue strategy is designed to allow quick and easy access to information at any point in the interaction. Most functionalities that are available at the initial "greeting" can be accessed at any turn in the dialogue.

The information contents in the responses of the system play an important role in the dialogue flow and hence users' perception of the system. W99 is equipped with short and informative prompts. Its dialogue strategy allows information to be repeated upon request. It can also accommodate for ambiguous and recurrent requests as illustrated in the following sub-dialogue:

User: Is there going to be a discussion on robustness?
W99: Would you like information about registration, transportation, technical and social activities or paper submission?
User: Robustness session please
W99: I do not understand your request. Please say registration, transportation, hotel, social, technical or paper submission.
User: OK, technical
W99: There will be six technical sessions ranging from speech ...
User: Technical please
W99: I do not have further information regarding this request. Please refer to our web page.

Robustness: In dialogue design, robustness is defined as the ability to maintain natural and constructive spoken dialogue irrespective of recognition errors. With the word recognition performance of fluent spontaneous speech ranging between 50-70%, it is not too surprising that dialogue strategies that are designed from text-input alone are likely to fail when operated from voice.

In W99, the dialogue strategy is generally designed to be user-initiative. In the event of key-phrase misrecognition, the dialogue switches to a system-initiative mode. This is identified when a key-phrase is either missing or has a low confidence score. For example, one simple strategy for activating the concept PAPER_STATUS is by recognizing paper and status. Should status be misrecognized, the dialogue would evolve as follows:

---

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

Recog: Need to know six us paper
W99: Would you like to know about the call for papers?
Recog: No 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. AUTOMATIC SPEECH RECOGNITION

The lack of data collection or field trials for conference registration and information access provides several challenges to acoustic and language modeling.

4.1. Acoustic Modeling

Due to the unavailability of acoustic data, Phase 0 (July'99 deployment) of the W99 system included off-the-shelf acoustic models from the *How May I Help You* (HMIHY) study [5]. These models included 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. The HMMs have been trained using maximum likelihood estimation (MLE) followed by minimum classification error (MCE) training [5].

During the period leading to Phase 1 deployment of W99 (August'99), a small corpus of 750 utterances was collected while the system was in operation, and was used for generating a new set of HMMs. These HMMs were trained by adapting Phase 0 models using MCE.

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

The challenge in building language models for the W99 system is providing different users the flexibility to speak freely. Without sufficiently large data collection, the diverse and unpredictable set of responses that we experience make W99 a challenge for language modeling.

In our initial effort in building language models for W99, a stochastic word bigram was created using the HMIHY field-trial data – a rather different application than W99. This model was continuously adapted using text data that were collected from our web-based dialogue system. 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 and Phase 1. 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. 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 2000 sentences and a lexicon of 1400 words. Compound language models were also generated to accommodate for embedded digits in the dialogue.

5. SYSTEM EVALUATION

Evaluating spoken dialogue systems remains an open problem. Most systems tend to use the number of dialogue turns, duration of each interaction, as well as other factors to determine the dialogue success rate [3]. In this section, we present an experimental study for the evaluation of the W99 system.

Our study was conducted on 50 subjects, each being asked to perform four different tasks. These subjects had little or no knowledge of speech processing technology nor any familiarity or prior training of W99. Over 50% of them were non-native English speakers.

Each task included finding certain information about the workshop. Subjects were instructed to (a) speak naturally and fluently to W99, interrupting the system at any time during the dialogue, and (b) test system robustness by calling from various locations, and by providing W99 with information that may be wrong or irrelevant.

The four tasks are summarized as follows:

1. Register for the workshop using a preassigned (random) card number and find the cost of attendance.
2. Find hotel information, room rates and directions.
3. Obtain guidelines for writing a technical paper. Also find submission dates and where to send papers to.
4. Exchange a natural dialogue with W99 on anything relevant, or not, to the ASRU workshop.

The actual wordings used to instruct the subjects were carefully selected so that not to overlap with the key-phrases used in the NLU module. The four tasks were composed in a way that can evaluate the capabilities, limitations and robustness of the W99 system.

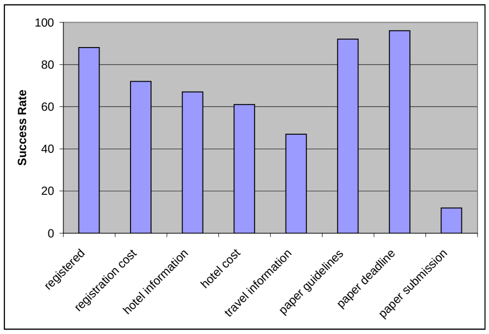

Figure 2. Success rates for W99 functionalities.

Subjects completed two sets of questions. The first included specific questions about system functionalities, and the second were oriented towards the subjects' interaction with W99. The results for the first set of questions are shown in Figure 2 which suggest the following: Over 88% of the subjects claimed to have registered successfully, but only 72% managed to obtain information about registration cost. 61% of the subjects obtained hotel information, and 67% found out the cost of the hotel room. Explicit directions to the hotel was not given by W99, but 47% of the subjects were satisfied to find only travel information. 92% of the subjects obtained guidelines for writing their papers, 96% found the deadline dates for paper submission, but only 12% claimed to have figured out where to submit technical papers to (electronic submission). The latter functionality was not supported by W99, hence the low score.

The second set of questions was asked after each task and included the following:

1. Did W99 understand what you said?
2. Was it easy to find the information you wanted?
3. When the system was unable to give you the information you wanted, were its responses sensible?
4. How would you rate your overall impression and interaction with W99?.

Scoring was tabulated from 1-5 with 1 being *almost never, very difficult,* or *very bad* and 5 being *almost always, very easy* or *very good.*

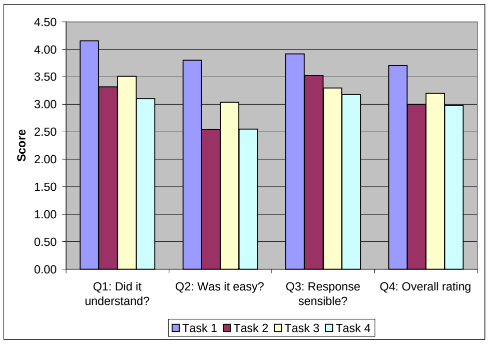

Figure 3. Scores for Subjects' interaction with W99.

The responses to the above four questions for all the four tasks are shown in Figure 3. The majority of the subjects were somewhat satisfied that the system "understood" them, giving Question 1 an average score of 3.5 across all tasks. The lowest score, with an average of 3.0, was given to Question 2, which relates to the ease-of-use of the system. An interesting result is an average of 3.5 being given to Question 3. This is an important test of system robustness and reflects the ability of W99 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 task 4.

In the following, we present an evaluation of the W99 system both in terms of recognition performance and concept accuracy on 2095 utterances that were collected in the experiment. Recognition performance represents the word error rate including insertions, deletions and substitutions, while the concept accuracy reflects the discrepancy in the NLU output when using the recognized speech as opposed to the transcription. We have tested both Phase 0 and Phase 1 developments of the system. As pointed out earlier, new sets of acoustic and language models were deployed in Phase 1 which have been adapted on 750 "live" utterances that were recorded in July'99. The data was collected anonymously from various callers who were mainly testing and exploring the system capabilities.

	WER	**CA**
Phase 0 (Jul'99)	49.8	70.4
Phase 1 (Aug'99)	46.7	74.8

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

The performance of the W99 system is illustrated in Table 1. Both Phase 0 and Phase 1 systems run at eight times faster than real time on an SGI R-10000 machine. Although the word error rate is 46.7% for Phase 1 system, 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 note that for this operating point, users' overall rating of W99 was 3.2.

6. SUMMARY

This paper presented the W99 system, a spoken dialogue system that has been deployed for the 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.

An experimental study was reported in this paper in which 50 subjects were asked to exchange a natural dialogue with W99 on four different tasks. Scoring from 1-5, W99 achieved an overall rating of 3.2, which corresponded to a word error rate of 46.7% 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. 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 these ratings as well as others that we have reported in this paper are a clear indication of a successful application using spoken dialogue systems.

Acknowledgments

The authors would like to thank V. Goffin, R. Knag, M. Beutnagel, and I. Zeljkovic for their technical contribution. They also like to thank D. Berkley and L. Rabiner for their support.

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] 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.

[4] 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.

[5] 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.

[6] 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.