VPQ: A SPOKEN LANGUAGE INTERFACE TO LARGE SCALE DIRECTORY INFORMATION

B. Buntschuh, C. Kamm, G. Di Fabbrizio, A. Abella, M. Mohri, S. Narayanan, I. Zeljkovic, R. D. Sharp, J.Wright, S. Marcus[#], J. Shaffer[*], R. Duncan[*] and J. G. Wilpon

AT&T Labs - Research
180 Park Ave., Bldg. 103
Florham Park, NJ 07932
Email: bb@research.att.com

ABSTRACT

This paper describes VPQ (Voice Post Query), a dialog system that provides spoken access to the information in the AT&T corporate personnel database (>120,000 entries). An explicit design goal is to have the user's initial interaction with the system be rather unconstrained and to rely on tighter, prompt constrained, dialog only when absolutely necessary. The purpose of VPQ is both a) to explore and exploit the capabilities of "state of the art" speech recognition systems for this high-perplexity task, and b) to develop the natural language understanding and dialog control components necessary for effective and efficient user interactions. The VPQ task spans a wide range of possible dialog scenarios. They range from simple "one-shot" to complex multi-turn interactions. The former correspond to interactions where the initial utterance is unambiguous and the system's response appropriately terminates the interaction either by providing the desired information or completing a call to the requested person. The more complex interactions occur primarily whenever ambiguities or errors require resolution. Current speech recognition accuracy of 80% is adequate to pursue such an ambitious task. This paper highlights the inherent challenges in such a task, the major components of the system, the rationale for their design, and how they perform. The VPQ project targets a variety of access devices, including telephony, desktop and handheld devices offering multi-modal user interfaces. In this paper we focus on describing the telephony interface.

1. INTRODUCTION

Advances in automatic speech recognition (ASR) algorithms and increasingly more powerful computers are opening up new arenas for spoken language applications. From the telecom service provider's perspective, automating directory assistance (DA) is one of the more compelling of these opportunities. The number of entries in large corporate directories approaches that of directories of medium-sized cities. Providing an easy to use and effective spoken dialog interface to such corporate information is a challenging application for developing and testing the technology needed for automated large-scale directory access. In an effort to explore our current capabilities in these areas, we are developing a spoken interface to the information in the AT&T Post corporate database, which has over 120,000 entries. Post information has traditionally been retrieved either via a command line or web based text interface called PQ (Post Query). The VPQ project (Voice Post Query) provides an interactive speech interface to this information as well as optional call completion. The user can query for various attributes contained in the Post database, request a specific action to be taken, or simply speak a name, in which case the default action is that of a voice-activated dialer.

In addition to the challenge of accurately recognizing speech in such a large domain, the inherent ambiguities in the task require effective dialog management in order to avoid a burdensome user interface. A major assumption for this first phase of the project is that there is no a priori information as to the likelihood of a particular entry being requested, i.e. there is no training corpus for language or dialog models.

The first section of this paper describes the task and its complexity. The second describes the platform components and their performance. The last section reviews the status of the entire system and the ongoing efforts to improve it.

2. VPQ: TASK COMPLEXITY

The electronic corporate employee directory that is used in this project is, like most directories, both constantly changing and error prone. In order to be able to characterize the task, we took a snapshot of the data, processed it and reorganized and augmented it to facilitate our needs. An ongoing challenge is how best to incorporate the daily changes to the data, especially since all but the most sensitive information can be modified by the employees themselves. Although this is more of a pure database issue, it has ramifications to this project because people often ask for employees that have moved on. We believe it better to retain some information about these employees and provide their status as feedback, rather than have the system deal with what will likely be out of vocabulary input.

The first directory processing task is to generate the list of names that will be used for the speech recognition grammars. While this may seem trivial, the data are far from pristine, and a fair amount of effort is required to transform the information into a usable canonical form. The preprocessor generated entries corresponding to a) surname, b) first name followed by surname, and c) preferred name followed by surname (if a preferred name was specified and differed from the first name). In addition, heuristics were applied to automatically determine additional variations that were assumed to be reasonable ways to refer to the individual. For example, an entry with an initial in the first name field and a name in the middle name field generated an additional entry for middle name followed by last name. As a result, a directory entry with first name R., middle name Douglas, preferred name Doug, and last name Sharp generated entries of Sharp, Douglas Sharp and Doug Sharp, but not R. Sharp.

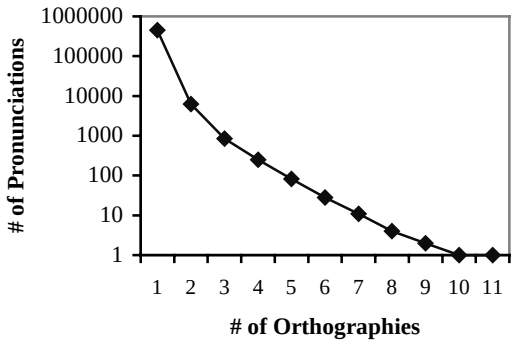
The snapshot of the employee directory that was used to generate the lexicon and grammars for off-line testing of the

VPQ system contained listings for 122,908 individuals. The preprocessing step generated 13,973 unique first and preferred names, 44,772 unique surnames, and 130,041 unique full names (first-last name or preferred–last name). The name-list that is used in the top-level grammar has 174,772 entries, allowing for either full names or just surnames.

The dialog manager dynamically adapts to the current conditions to expeditiously resolve ambiguities, uncertainties and error conditions. The principal ambiguities that characterize this task are both *homophonic*, i.e. a particular utterance may correspond to many distinct orthographies, and *multiplicative*, i.e. there can be multiple entries that match a particular orthography.

Figure 1 illustrates the homophonic complexity of the task. While the vast majority of pronunciations in the database are unique to a single orthography, a single pronunciation may represent as many as 11 different orthographies. In this case, the pronunciation /kin/ was generated for the orthographies Keane, Keene, Kiehn, Keen, Kuehn, Kuehne, Keehn, Kean, Kiene, Kyhn, and Keihn.

Figure. 1 Homophonic Complexity



However, even when the homophonic ambiguity is resolved, the pronunciation often maps to more than one individual, requiring further disambiguation to find the listing that the user wants. The most frequent surname is Smith, with 1080 entries. There are also 4 employees with the surname Smyth. A spoken name pronounced as /smIθ/ can map to either of these orthographies. The most frequently occurring full name is James Smith, with 21 entries. There are 623 full names that occur four or more times in the directory.

3. SYSTEM COMPONENTS

This section describes the major components of the system - speech recognizer with associated grammar utilities, database, natural language processor, dialog manager, and the underlying computer telephony infrastructure.

3.1 Automatic Speech Recognition

AT&T's Watson speech recognition technology [4,7] is used to provide high accuracy, speaker independent, speech to text in a telephony environment. The speech engine is capable of running grammars with vocabularies exceeding one million words in real-time on a standard RISC processor. While the core technology in Watson has been proven on several DARPA tasks [5], the challenge of very large vocabulary recognition in a task where a training corpus was not available has led to some new algorithms for automatic weight assignment in finite state networks.

The system does not currently take advantage of N-best or lattice result capabilities of the Watson ASR engine, although it does make use of utterance verification confidence measures. In order to determine whether portions of the recognized utterance correspond to out-of-vocabulary words or incorrectly decoded in-vocabulary words, a likelihood ratio based hypothesis testing is used to provide confidence measures for the decoded words [3]. The alternate hypothesis is obtained in a second pass through the decoder using HMMs that are trained from a development set comprising utterances representative of the VPQ task. These are distinct from HMMs used in the first pass which were trained using the task independent data. The scores obtained are passed along to the natural language and dialog modules.

3.1.1 VPQ Grammars. The active grammar for the first interaction with the user has a manually specified command and control syntax "surrounding" the name-list described above. It allows for a variety of ways to request a) most attributes associated with an employee such as phone, fax or mobile number, e-mail address, location, etc., b) help, and c) actions such as calling, paging or faxing. In the surround grammar, the name-list non-terminal is expanded four times, once in a possessive form. None of the grammars used in this phase of the project employed N-grams or back-offs, although this is planned once we collect enough data. The pronunciations used were derived from the frontend of the Watson TTS engine. There are over 200,000 distinct entries in the lexicon.

As an initial test for coverage of the surround grammar, 115 volunteer subjects were given directory query task scenarios and were asked to call in to a mock-up of the system to request the information specified in the scenario. When subjects called the system they heard the open-ended prompt "VPQ. How may I help you?" Each caller's responses were recorded and transcribed.

For example, the scenario "You have a meeting with Mike Armstrong, but don't know at which AT&T location this person works" elicited responses that included:

"What's Mike Armstrong's room number?"
"Please give me the location for Mike Armstrong".
"I'm looking for Mike Armstrong's location."

Over 550 recordings of responses to the initial prompt were obtained. Only about 30% of the requests were completely within the original surround grammar. Among the more frequent deviations from the initial grammar were phrases like "provide me with", "look up", "find", "contact information", "looking for" and "listing". In addition, 19% of the requests were contained the word "please". The requested name was spelled as well as spoken in 11.5% of the requests. Requests for multiple database attributes (e.g., "Can you tell me the phone number and room number for Mike Armstrong?") were also observed. The results of this experiment were used to augment the surround grammar to include what were originally the more frequent out of grammar fragments.

The ASR grammar represents proper names in a phonetic form delimited by lexical semantic tags rather than standard orthographic form. The recognized phonetic form of a name is easily isolated in parsing and can then be efficiently searched for in the database to retrieve all candidate entries, essentially retrieving an equivalence class for that pronunciation.

3.1.2 ASR Performance. To estimate ASR performance for this task, web-based prompted data collection provided a corpus of spoken full, last and spelled last names that were randomly chosen from the corporate directory. ASR results, presented in Table 1, are for real-time performance using grammars corresponding to a 2k fullname, 174k fullname and the top-level grammar, where the 174k entry name-list occurred in four places. The grammar state and arc totals are for fully optimized grammars before composition with the lexicon[4]. The test set for the results presented consisted of 520 spoken fullnames that were all in the grammar.

Task	Grammar States/Arcs	Word Error Rate
2k fullnames	1633/6446	3%
130k full names	21789/292588	15%
174k full & last names	22599/378817	16%
174k w/surround	44464/761262	20%

Table 1. Real-time ASR Performance

In addition to word-based grammars, the VPQ application uses spelling as an important disambiguation method. Even though we expect to create spelling grammars dynamically based on dialog context, we tested performance for spoken spelled surnames for the entire AT&T population in the event that conditions required such an interface. The spelling grammar currently does not relax the input to allow for colloquial spelling such as described in [6] (e.g. "E as in Edward"), but the data collected in the prompted scenario showed that most people used straightforward spelling. Table 2 presents ASR test results for grammars corresponding to 2000 and 55230 spelled surnames using specialized alpha-digit acoustic models. These results demonstrate that spelling could be an effective solution.

Number names in grammar	Word(letter) error rate (%)	String (names) error rate (%)
2,000	0.6	1.8
55,230	1.9	7.0

Table 2. Spoken Spelled Surname Performance

3.2 Database

Since we needed to augment the directory information for this task, we decided to populate our own LDAP directory. The directory schema has been designed to efficiently accommodate the inherent ambiguities in the task. The entry for each unique person has as attributes canonical corporate data. Associated with each such entry is a group of "VPQname" entries, each of which corresponds to a unique orthographic representation for that entry. For instance "Lawrence Rabiner" has three VPQnames, "Rabiner", "Lawrence Rabiner", and "Larry Rabiner", each of which is derived from the official corporate database. Each of these orthographies may have multiple phonetic representations, which results in greater than 200,000 distinct entries in the lexicon.

Recognizing the propensity for homophonic ambiguities amongst the names in the directory, we create the ASR grammars with a phone level representation of the names. This does not effect the accuracy of the recognition, and, combined with the fact that the database is indexed on the phone level representation of names, facilitates retrieval of all entries that correspond to a specific utterance. This approach results in minimal database interactions when an ambiguous phrase is spoken, and minimal overhead otherwise.

Rather than group all of the orthographic and phonetic information in the same entry as the rote corporate data, we created a separate VPQname object-class that maintains the correlation between the separate orthographies and phonetic representations. We chose to do this both to facilitate the updating of information and to be able to provide feedback to the user consistent with what was spoken. For instance, if someone requested information about "Jim Smith", the system does not need to respond with the entry's canonical given name "James", but can respond with precisely the form that the user spoke.

In addition, we are developing an automated directory update procedure that incorporates modifications in the corporate directory as well as allowing for user-specified input, such as personalized pronunciations.

3.3 Natural Language and Dialog Control

The natural language (NL) and dialog manager components have been successfully used in other spoken language tasks [1,9,10]. Each has a reusable algorithmic base that requires only a small amount of task specific programming. The natural language processor uses a combination of lexical semantic tags present in the ASR output and robust topic classification to produce canonicalized data for the dialog manager. Since this module was designed for use with fully stochastic grammars, the presentation of results to the dialog manager will not need to change once we gather enough data to train and use such language models.

The dialog manager [1] employs dynamic constraint analysis to resolve ambiguities in the most effective fashion given a particular cost function. For VPQ, this amounts to minimizing the number of turns by requesting information that is likely to be known by the user. The dialog manager is responsible for determining the structure of the interaction with the user based on the current context of the interaction and the most recent ASR/NL result. A collection of dialog strategies is applied to the current context to determine the next system action. In contrast to systems that prompt for a set of predetermined attributes, this system will behave differently depending on the recognition results of the previous turn.

Two examples of disambiguation sub-dialogs are shown below

System: VPQ, how may I help you?
User: Call Marolla.
System: I have listings for Paul Marolla and Janet Marolla, which one do you want?
User: Janet Marolla**.**
System: Calling Janet Marolla**.**

In this case, there are only two listings that match in the database, and they differ in the first name attribute. This attribute is one of a small set of the database attributes that have been prespecified as likely to be known by the user. Since the set of alternatives is small (less then three), the dialog manager explicitly mentions each alternative and asks the user to select the correct one.

> **System:** VPQ, how may I help you?
> **User:** What is John Smith's room number?
> **System:** I have 21 listings for John Smith. Do you know the work location for this person?
> **User:** Middletown.
> **System:** The room number for John Smith in Middletown is MT 2H-314.

In this case, there are 19 individuals named John Smith, one named Jon Smith, and one named John Smyth. However, rather than asking for spelling of either the first or last name, the dialog manager first asks for location information because it is a more efficient attribute for disambiguating the listing.

The dialog manager also controls error handling for conditions of ASR time-out, uncertainty and misrecognition. Uncertainties are detected in the utterance verification process and passed on to the natural language component where they are incorporated in the classification process. Even with very robust ASR, the confusability of the top-level grammar leads to many errors that will be hard to avoid. For instance, there are the common substitutions of "page" for "Dave, and "call" for "Paul". Since the dialog manager implicitly confirms what was recognized, pure ASR errors generally elicit negative responses from the user which are detected in the next dialog turn as very low confidence utterances by the recognizer. In the case of the substitution of "call X" for "Paul X", if the entry for X is unambiguous, then the system actually takes the appropriate action, despite the error. We rely on the user being able to utter a cancel command during the confirmation prompt. In previous telephony agent applications, users much preferred this method rather than being burdened for an explicit confirmation [2].

3.4 Computer Telephony Platform

The entire system is built upon a robust, ECTF standards based computer telephony infrastructure. The platform supports multi-channel applications running ASR with dynamically creatable grammars, barge-in, TTS, recorded prompts, DTMF detection and telephone call control. The platform makes extensive use of client/server paradigm in order to distribute the processing demands, especially of the ASR and database components. We have developed an abstraction layer that simplifies the interface to the underlying complexity of the CTI infrastructure while merging the other components into a common application fabric. The platform is fundamentally asynchronous. However, to simplify the dialog design, many components, such as database access, are currently implemented in synchronous fashion since they generally complete rapidly.

4. STATUS AND CONCLUSIONS

The components of VPQ described in this paper have recently been integrated into a working system. A live trial for the AT&T employee population is forthcoming. This trial will use the full surround grammar syntax with the larger (174,000) name-list. We realize that this fairly open-ended grammar, although not truly unconstrained, is ambitious. All spoken utterances will be collected and will contribute to future training corpora for the development of stochastic grammars and improved acoustic models. Until this data becomes available, we rely on utterance verification to gracefully accommodate out of grammar inputs. Experience with the system will determine the appropriate tradeoff between input flexibility and performance. Many aspects of the system draw from the knowledge gained from other disparate tasks [1,2]. Our experience shows that we can expect poor performance initially unless the users are made explicitly aware of the capabilities of the system. To this end we provide a web-based training tutorial, as well as the complete help that is available via the phone.

REFERENCES

1. Abella, A. and Gorin, A. L. "Generating semantically consistent inputs to a dialog manager." *Proc EuroSpeech '97*, 1879-1882, 1997.
2. Kamm, C., Narayanan, S., Dutton, D. and Ritenour, R. "Evaluating spoken language systems for telecommunications services". Proc. Eurospeech '97, 2203-2206, 1997.
3. Lleida, L. and Rose, R. "Utterance verification in continuous speech recognition: decoding and training procedures." *ICASSP '96*, 507-510, 1996.
4. Mohri, M. and Riley, M. "Weighted determinization and minimization for large vocabulary speech recognition." *Proc. EuroSpeech '97*, 131-134, 1997.
5. Mohri, M., Riley, M., Hindle, D., Ljolje, A. and Pereira, F. "Full expansion of context-dependent networks in large vocabulary speech recognition." Proc. ICASSP '98, 665-668, 1998.
6. Seide, F. and Kellner, A. "Towards an automated directory information system." *Proc. EuroSpeech '97*, 1327-1330, 1997.
7. Sharp, R. D., Bocchieri, E., Castillo, C., Parthasarathy, S., Rath, C., Riley, M. and Rowland, J. "The Watson speech recognition engine." *Proc. ICASSP 97*, 4065-4068, 1997.
8. Whittaker, S. J. and Attwater, D. J. "The design of complex telephony applications using large vocabulary speech technology." *Proc ICSLP 96*, 705-709, 1996.
9. Wright, J. H., Gorin, A. L. and Abella, A., "Spoken language understanding within dialogs using a graphical model of task structure." *Proc. ICSLP 98*, 1998.
10. Wright, J.H., Gorin, A. L., and Riccardi, G., "Automatic acquisition of salient grammar fragments for call-type classification." *Proc. EuroSpeech '97*, 1419-1422, 1997.

† S. Marcus is now affiliated with International Asset Systems, LTD.

* J. Shaffer and R. Duncan are affiliated with Mississippi State University.