Exploiting Context Information in Spoken Dialogue Interaction with Mobile Devices*

Stephanie Seneff[1], Mark Adler[2], James R. Glass[1],
Brennan Sherry[1], Timothy J. Hazen[1], Chao Wang[1], and Tao Wu[2]

MIT Computer Science and Artificial Intelligence Laboratory[1]
Cambridge, MA 02139, USA

Nokia Research Center[2]
Cambridge, MA 02142, USA

Abstract. Today's mobile phone technology is rapidly evolving towards a personal information assistant model, with the traditional cell phone morphing into a networked mobile device that is capable of managing many other personal needs beyond communication with other humans, such as a personal calendar, address book, photography, music, etc. As the mobile phone becomes more complex, it is becoming increasingly difficult for users to interact effectively with the various services that are offered. Our vision is that natural spoken conversation with the device can eventually become the preferred mode for managing its services. It is our belief that contextual information (dialogue context, personalization to the user, situation context, and the graphical interface) will play a critical role in the successful implementation of such a model. This paper focuses specifically on the many ways in which contextual information can be used to improve the effectiveness of spoken dialogue. We describe two applications currently under development, a calendar management system and a photo annotation and retrieval system.

1 Introduction

The mobile phone is rapidly transforming itself from a telecommunication device into a multi-faceted information manager that supports both communication among people as well as the manipulation of an increasingly diverse set of data types that are stored both locally and remotely. As the primary utility of the device moves beyond supporting spoken human-human communication, speech and language technology will prove to be vital enablers for graceful human-device interaction, and for effective access to rich digital content.

The need for better human-device interaction is clear. As personal devices continue to shrink in size yet expand their capabilities, the conventional GUI model will become increasingly cumbersome to use. What is needed therefore is a new generation of user interface architecture designed for the mobile information

* This research is supported by the Nokia-MIT Alliance.

device of the future. Speech and language technologies, perhaps in concert with pen-based gesture and conventional GUI, will allow users to communicate via one of the most natural, flexible, and efficient modalities that ever existed - their own voice. A voice-based interface will work seamlessly with small devices, and will allow users to easily invoke local applications or access remote information.

This paper describes our vision for systems that enable humans to communicate with their mobile phones via natural spoken dialogue, in order to exploit intuitive interaction modes to access available services. Many potential applications can be imagined, ranging from the relatively simple task of address book or calendar management to the much more complex task of photo annotation and retrieval. For this vision to be realized, many technology advances must be made: we must go beyond speech recognition/synthesis, and include language understanding/generation and dialogue modeling.

A central theme critical to the success of the system is the use of contextual information, not only to solve the tasks presented to the system by the user, but also to enhance the system performance in the communication task. Context-awareness has become an important issue in human-computer interaction (HCI) [10] and in computer-supported cooperative work (CSCW) [13], as advances in technology move the site and style of interaction beyond the desktop and into the larger real world [19]. In the conventional GUI setting, context can simply be the status of the application itself in which a user's mouse/keyboard actions are interpreted. However, with today's mobile devices, context can include any information about the environment/situation that is relevant to the interaction between a user and an application [4]. It has become a shared idea in the pervasive computing community that enabling devices and applications to adapt to the physical and electronic environment would lead to enhanced user experience [4]. Some application scenarios are illustrated in ringing phone and medical alarm systems [2] and a highly sophisticated conference assistant [4].

In this paper, we demonstrate the usage of context in two speech-based mobile phone applications: a calendar management system and a photo annotation and retrieval system. We have attempted to organize "context" into four broad categories: (1) dialogue context, (2) personalization to the user, (3) situation context, and (4) graphical interface. In the following sections, we will first describe what we mean by each of these context categories, giving examples where appropriate. We will then describe in more detail the two systems that we are developing, where we have begun to exploit some of the envisioned context conditions. We conclude with a summary and a discussion of our future plans.

2 Contextual Influences

In this section, we discuss the four distinct dimensions of context which we believe are important for optimizing the user's experience in interacting with spoken dialogue systems deployed for mobile phones.

2.1 Dialogue Context

Perhaps the most crucial, and potentially most difficult, contextual influences are those that are provided by the immediately preceding dialogue. The fact that goals are achieved in our model via a *conversation* is a distinguishing feature compared with the notion of verbal *commands* spoken in isolation. To give a concrete example, suppose the user asks, “What meetings do I have today?” The system responds with, “You have a 10 a.m. meeting with Sally and a 2 p.m. meeting with Allen.” The user then says, “Could you add a meeting with Karen just after my meeting with Allen?” The system would appear to be quite inept if it were to respond with: “You have two meetings with Allen, one today at 2 p.m. and the other on Friday at 11 a.m. Which one do you mean?” Clearly, it needs to keep track of salient reference objects (e.g., the meeting with Allen at 2 p.m. today), which may have been introduced into the dialogue either by the user or by the system.

It is not always clear when to inherit constraints from the dialogue context and when to leave them out, and part of the challenge is to develop proper heuristics to allow the system to make reasonable decisions regarding context, as well as keeping the user well informed of these assumptions. One approach we are exploring is to assume inheritance somewhat aggressively, but to use implicit confirmation to inform the user of our assumption. Thus, to continue with our example, if the user, after adding the meeting with Karen, says, “I’d like to schedule a meeting with Ellen and Carl,” the system would respond with “At what time today would you like to schedule the meeting?” The user can then say, “Tomorrow at 9 a.m.” to overrule the inappropriately inherited date, while also providing the time. This last sentence is an example of a sentence fragment in response to a system prompt, and of course the system must also keep in focus across dialogue turns the fact that the meeting is with Ellen and Carl.

2.2 Personalization to the User

In prior research, we have developed spoken dialogue systems that support telephone access to information sources such as weather [18] and flight information [14]. For these systems, users can be calling a remote system from any telephone, and, even if caller-id were to be exploited, there is no guarantee that the same person calls from a particular landline phone over multiple calls. Furthermore, it is likely that most users would call too infrequently to allow an adequate model of their usage patterns to emerge.

In contrast to this model, a mobile phone is usually a personal device carried by a single user. One of the enormous advantages of working with a system that is intended to serve only one user is the opportunity this offers to personalize the system to that particular user. Hence, we can expect to be able to monitor many aspects of their behavior/data and adjust system parameters to better reflect their needs.

There exists already an extensive body of research on the topic of speaker adaptation of the acoustic models of a speech recognizer [6, 8, 9]. We can envision a system which in the early phases *adapts* generic acoustic models, eventually switching over to *speaker-trained* models once enough data have been captured.

Beyond acoustic models, we should also be able to train the language models to favor patterns the user has grown accustomed to. This includes both the particular ways the user typically expresses certain requests, as well as the general topics the user is interested in. For example, one user may be very inclined towards interacting with the music player whereas another may make frequent inquiries about the calendar or ask for reminders.

Personal data that are relevant to the application domain can also be used to help in the speech recognition task. For example, it would be logical for a calendar application to assume that people already in the user's address book are likely candidates for a future meeting. Thus, the recognizer should automatically augment its name vocabulary to include everyone in the address book. This is currently being incorporated in our system by using the address book names using the dynamic vocabulary feature of Galaxy. This can be further extended to include external sources as well, either by connecting to a corporate directory, or in a meeting environment to include the attendees, since the meeting could provide context for future meetings. A further variation would be to extend this another level to include all the attendees' address book information as well, depending on access policies.

2.3 Situation Context

We use *situation context* to refer to the environment that the application is in[1]. The most relevant factors for our applications include geographical locations and time.

For example, an interesting issue for calendar management is the reference time zone, which would naturally be omitted in conversation. If the user is on vacation in Paris, should it assume they will be making references to time relative to Paris' time zone, or should it perhaps instead prompt them for clarification of their intended time zone at the onset of a conversation? In global corporations with video and voice conference calls, time zone clarification is essential, and usually specified. However, with global travel, should the system assume that the time is local to the location of the user at the time of the meeting? In that case, the user's travel itinerary provides essential information for scheduling.

The photo annotation and retrieval application has many interesting opportunities to exploit metadata associated with photographs. For example, it would likely be beneficial to organize photographs around place (GPS or cell tower location) and time, metadata that are now typically automatically associated with a photograph when it's first captured. A series of photographs taken in Hawaii over a period of a week could be automatically organized into a cluster group. The system could recognize that the location is Hawaii directly from the GPS code. The user could also be prompted to provide a verbal description of the

[1] This roughly corresponds to the notion of context used by the pervasive computing community [4].

cluster, such as "These are pictures from our honeymoon." Individual photos would then inherit group properties, including the audio index, such that the user could say, "Show me all the pictures of the fishing trip on our honeymoon," – the group match to "honeymoon" would be automatically joined with the individual matches to "fishing trip" in the retrieval step. Again information from the user's existing calendar and travel itinerary could be used to extend the vocabulary to include typical tourist sites in the destination locations. A history of a user's behavior and preferences could be used to help guide the vocabulary augmentation by including, for example, names of local vineyards for wine connoisseurs who would likely add a side trip for wine tasting.

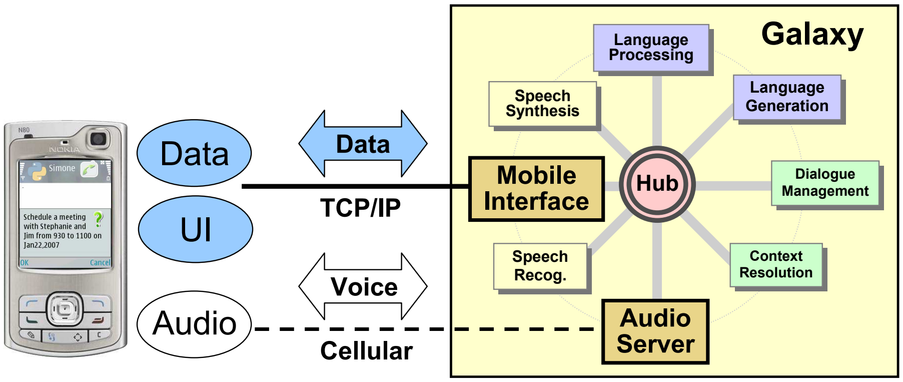
2.4 Graphical Interface

When telephone-based applications were first developed, the assumption was made that there would be no graphical display available, and that the telephone keypad could at best be utilized for simple tasks such as an "abort" signal or, with limited success, to correct misrecognitions such as a date encoded as a four digit string [16]. In recent years, remarkable progress has been made in the resolution of the display screen of a mobile phone, such that it becomes potentially a much more significant player in the dialogue interaction, both in terms of delivering information to the user visually and in terms of providing an opportunity for multimodal inputs. However, it remains a research challenge to understand how to most effectively utilize the display during spoken dialogue. Certainly the opportunity to select among multiple competing candidates produced by the recognizer provides a powerful mechanism to lessen the impact of recognition errors. At the critical point when the system is about to modify a database, such as the calendar, particularly if it involves a deletion, it could be preferable to *require* that confirmation be entered at the keypad rather than verbally, since one could run the risk of a misrecognition of a spoken confirmation.

The photo retrieval application would benefit significantly from a much larger screen than a mobile phone can provide. Thus it is interesting to consider a configuration in which the phone serves as an I/O device for a desktop computer, where the user can display multiple photographs simultaneously. For such a configuration to be effective, the dialogue manager would need to have a dual-mode design, based on the size of the available display screen. With a small screen, it could provide a summative verbal description of a collection of photos in focus, asking the user to narrow down to a specific one before providing an image. A large screen invites greater opportunities, for example, selecting a subset of the photos being displayed, and asking the computer to perform some operation on the selected subset, such as "email these pictures to my sister."

3 Current Status

In our experiments, we have developed a common architecture for both calendar management and photo annotation applications. As shown in Figure 1, this



Schedule a meeting with Stephanie and Jim from 930 to 1100 on Jan22,2007

Fig. 1. System Architecture: the GALAXY system runs on a remote server, where the various technology components communicate with one another via a central hub. The mobile phone utilizes the voice channel for speech I/O and the data channel to transmit information needed by the GALAXY system to interpret queries.

architecture consists of a GALAXY server [15] and a mobile device, connected with each other using voice and data channels. The user interacts with the client application on the device (i.e., the calendar or photo annotation application), primarily via spoken conversation. Speech is transmitted full duplex over the standard phone line. Application data, such as calendar entries stored on the phone or responses from GALAXY, are transmitted via the data channel.

Although we have not yet addressed many of the issues that are discussed above, we have developed prototypes of the calendar management and photo annotation applications with an emphasis on natural dialogue interaction. We are currently launching a data collection effort in both domains, which will guide and inspire future enhancements to these applications. In the following, we describe these two systems in somewhat more detail.

3.1 Calendar Management Domain

The calendar application enables a user to manage the calendar on mobile phones via verbal interactions in English or Chinese. A user can query, add, delete, or change calendar entries by speaking with a dialogue system. The system can engage in sub-dialogues with the user to resolve ambiguities or conflicts, or to obtain confirmations. The system capabilities are best illustrated via an example dialogue, as shown in Figure 2.

The application is built on human language technologies developed at MIT. Speech recognition is performed by the SUMMIT landmark-based system using finite state transducer technology [7]. We utilize the TINA natural language understanding system [17] to select the appropriate recognizer hypothesis and interpret the user's query. A context free grammar based on syntactic structure

S: Welcome to the Nokia telephone management system.
U: What meetings do I have today?
S: There are two meetings on January 17. The meeting with Jim between 10am and 11am, and the one with Mark between 2pm and 3pm.
(System resolves the date for "today," Phone displays meeting details.)
U: Cancel the first meeting.
S: Should I cancel the meeting with Jim on Wednesday January 17 between 10am and 11am.
(Phone displays details for confirmation.)
U: Yes. (Alternatively the user can press a button.)
S: Okay, I have erased the meeting with Jim.
U: Move my meeting with Mark to tomorrow at the same time.
S: That conflicts with your scheduled meeting with Jamey.
(Phone displays the details of the meeting in conflict.)
U: Put it before my meeting with Jamey.
S: Should I change the meeting with Mark on Wednesday January 17 to 2pm on January 18?
(Phone displays details of the proposed change.)
U: Yes.
S: Okay, I have changed the meeting with Mark.
U: Please schedule a meeting with Stephanie and Jim.
S: At what time should I schedule a meeting with Stephanie and Jim on January 18?
(System assumes the same date mentioned mostly recently.)
U: Next Monday from 9:30 to 11.
(User implicitly corrects the system.)
S: Should I schedule a meeting with Stephanie and Jim next Monday January 22 from 9:30am to 11am?
(System assumes morning time. Phone displays details for confirmation.)
U: Yes.
S: Okay, I have scheduled a meeting with Stephanie and Jim.
U: Thank you. Good bye.

Fig. 2. Example spoken dialogue between an expert user and the calendar management system.

parses each query, and transforms it into a hierarchical meaning representation that is passed to the discourse module [5], which augments the meaning representation with discourse history. A generic dialogue manager [12] processes the interpreted query according to an ordered set of rules, and prepares a response *frame* encoding the system's reply. Domain-specific information is provided to the dialogue manager in an external configuration file. The response frame produced by the dialogue manager is converted into a *response string* via the Genesis language generation system [1, 3]. Finally, a speech synthesizer prepares a response waveform which is shipped across the standard audio channel for the system's turn in the dialogue. In addition, a graphical display illustrates changes made to the calendar at the point where a proposed change to the calendar is being confirmed. We have developed two versions of this system, one supporting interaction in English, and the other one in Mandarin Chinese. We have set up a web-based system to perform a user study to assess the coverage of our calendar dialogues.

3.2 Photo Annotation and Retrieval

Our photo annotation and retrieval system is currently comprised of two components, one for annotating photos and one for retrieving. During the annotation stage, users can provide personalized verbal annotations of their own photographs. For maximum flexibility, the user is not restricted in any way when they provide these free-form annotations. As an example, the photograph in Figure 3 could be annotated by the user as:

Julia and Pluto at Disney World.

User annotations are automatically transcribed by a general purpose large vocabulary speech recognition system [7]. The current system utilizes a 38k word vocabulary. To compensate for potential misrecognitions, alternate hypotheses can be generated by the recognition system, either through an N-best list or a word lattice. The resulting recognition hypotheses are then indexed in a term database for future look-up by the retrieval process. In addition to the spoken annotation, metadata associated with the photograph is also stored in the database. This metadata can include various pieces of information such as the owner of the photograph, the date and time the photograph was taken, and the GPS location of the camera or digital device that took the photograph.

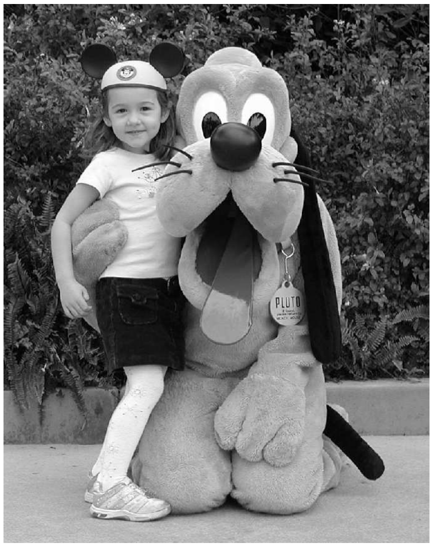


Fig. 3. Example photograph from the photo annotation system.

During the retrieval stage, the user can speak a verbal query to specify the attributes of the photographs they wish to retrieve. The system must handle queries that contain information related to either the metadata and/or the free-form annotations. For example, the user could refer back to the photo shown in Figure 3 with queries such as:

> *Show me John Doe's photo of Julia with Pluto at Disney World from December of 2005.*

This query specifies constraints on both the metadata of the photograph (i.e. whose photograph it is, and when it was taken) and the free-form information that describes the contents of the photograph.

To handle the photograph retrieval queries, such as the one shown above, we have developed a mixed-grammar approach. The speech recognition language model uses a constrained context-free grammar to handle the general carrier phrases associated with retrieval queries (i.e. "Show me photos of...") and the phrases for specifying metadata (i.e., "...taken in December of 2005."). To recognize the free-form annotation terms, a large vocabulary statistical n-gram language model is used. These are encoded as finite-state networks and combined within a single search network, which allows the system to transition appropriately among the component networks when analyzing the spoken utterance. For understanding, the grammar's output structure is converted directly into an XML representation of the interpreted query. The example query discussed above would generate the following interpretation:

```
<request>
   <search_terms> julia with pluto at disney world </search_terms>
   <month> 12 </month>
   <year> 2005 </year>
   <owner> John Doe </owner>
</request>
```

All of the functionality discussed above has been implemented within a client/server architecture in which photographs are first taken and annotated on a client camera phone. The photographs and audio annotations are sent over a data connection to the photo annotation server for transcription and indexing into a photo database. The photos can then be retrieved at a later time from the server via a spoken request from the phone. This audio request is processed by the photo retrieval server, and relevant photographs are returned to the user's client phone for browsing. We are running user studies to assess the current system.

We view this as an initial prototype that leverages traditional context information such as time and location with additional information that is more difficult to associate with photographs such as the subject in the image, as well as the source of the image, be it the photographer or which camera (or device) took the photograph.

One area that we view as critical for mobile retrieval is the ability to summarize a set of data by using context to partition the results [11]. For example,

```
U: Show me pictures from Disney World in December 2005.
S: There are 50 pictures from Disney World in December 2005; 20
   from Friday, December 23, 15 from Saturday, and 15 from Sunday.
U: Show me the ones from Christmas.
```

4 Summary and Future Work

In this paper, we have described two mobile phone applications, a calendar management system and a photo annotation and retrieval system, to exemplify the use of a speech interface as a natural means to control and communicate with mobile devices. We highlighted the many ways in which contextual information can be used to improve the effectiveness of spoken dialogue.

Although a remote server-based solution remains a reasonable approach, particularly when large databases such as photo repositories are involved, local processing is preferable in many circumstances. Thus, in addition to continuing the work described here, we are migrating from a server-based system towards a model where speech components reside on the mobile device. Our initial platform is the Nokia Linux Internet tablet, the N800, which provides a larger, touch sensitive screen. This will allow us to incorporate local context in a more straightforward manner (eliminating the requirement of connectivity to a remote server). We will initially focus on the speech recognizer and the Text-to-Speech (TTS) components, concentrating on two languages, Mandarin and English.

References

1. L. Baptist and S. Seneff, “Genesis-II: A versatile system for language generation in conversational system applications,” *Proc. ICSLP ’00*, V. III, 271–274, Beijing, China, Oct. 2000.
2. B. Brown and R. Randell, “Building a context sensitive telephone: some hopes and pitfalls for context sensitive computing,” *Computer Supported Cooperative Work*, 13:329–345, 2004.
3. B. Cowan. “PLUTO: A preprocessor for multilingual spoken language generation,” S.M. thesis, MIT Department of Electrical Engineering and Computer Science, February, 2004.
4. A. K. Dey, G. D. Abowd, and D. Salber, “A conceptual framework and a toolkit for supporting the rapid prototyping of context-aware applications,” *HCI Special Issue of Human-Computer Interaction*, 16:97–166, 2001.
5. E. Filisko and S. Seneff, “A Context Resolution Server for the Galaxy Conversational Systems,” *Proc. EUROSPEEECH-03*, 197–200, Geneva, Switzerland, 2003.
6. J. Gauvain and C. Lee, “Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains,” *IEEE Transactions on Speech and Audio Processing*, 2(2):291-298, 1994.
7. J. Glass, “A probabilistic framework for segment-based speech recognition,” *Computer Speech and Language*, 17:137-152, 2003.
8. T. Hazen, “A comparison of novel techniques for rapid speaker adaptation,” Speech Communication, 31(1):15-33, 2000.

9. C. Leggetter and P. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," *Computer Speech and Language*, 9(2):171-185, 1995.

10. T. P. Moran and P. Dourish, "Introduction to this special issue on context-aware computing," *HCI Special Issue of Human-Computer Interaction*, 16:87–95, 2001.

11. J. Polifroni, G. Chung, and S. Seneff, "Towards the automatic generation of mixed-initiative dialogue systems from web content," *Proc. Eurospeech*, Geneva, 2003.

12. J. Polifroni and G. Chung, "Promoting Portability in Dialogue Management," *Proc. ICSLP*, 2721–2724, Denver, Colorado, 2002.

13. A. Schimidt, T. Gross, and M. Billinghurst, "Introduction to special issue on context-aware computing in CSCW," *Computer Supported Cooperative Work*, 13:221–222, 2004.

14. S. Seneff, "Response planning and generation in the MERCURY flight reservation system," *Computer Speech and Language*, 16:283–312, 2002.

15. S. Seneff, E. Hurley, R. Lau, C. Pao, P. Schmid, and V. Zue, "GALAXY-II: A reference architecture for conversational system development," *ICSLP '98*, 931–934, Sydney, Australia, December, 1998.

16. S. Seneff and J. Polifroni, "Hypothesis selection and resolution in the MERCURY flight reservation system," *Proc. DARPA Human Language Technology Workshop*, San Diego, CA, pp. 145–152, 2001.

17. S. Seneff, "TINA: A natural language system for spoken language applications," *Computational Linguistics*, 18(1):61–86, 1992.

18. V. Zue, S. Seneff, J. Glass, J. Polifroni, C. Pao, T.J. Hazen, and L. Hetherington, "Jupiter: a telephone-based conversational interface for weather information," *IEEE Transactions on Speech and Audio Processing*, 8(1):85–96, 2000.

19. M. Weiser, "The computer for the 21^{st} century," *Scientific American*, 265:94–104, 2001.