

Fleximodal Conversational Interaction: Using Voice and SMS to Find Places

Abstract

Theories of conversation apply very well in this era of short messages and rapid exchange on mobile devices. In particular, conversational models for information-seeking dialogues have advanced near to the point of commercialization (Larsson and Villing 2007). Such models apply well to both speech and text messaging interaction on cellphones. By allowing the user to interact in short messages using whichever channel is most efficient at that point in the interaction, may provide a means for both improving the efficiency of search and relevance of search results -- as well as improving the experience of search on cellphones.

Lisa Harper

*Doctorate in Communications Design
Professor Kathryn Summers / IDIA 612
28 November 2007*

Fleximodal Conversational Interaction: Using Voice and SMS to Find Places

This paper provides a literature review for a research system under design. The system goal is to assist users with finding coffee shops using a cellphone. After reviewing several commercial systems and research literature in the area of conversational interaction, we believe that a mixed modal (speech and SMS) conversational interface can be leveraged to enhance both the *efficiency* of search and *relevance* of search results -- as well as to improve the user experience of search on cellphones. This paper provides an integrated literature review to support this hypothesis. It starts with a discussion of affordance of cellphones, networks, location-based services, and message formats. It then turn to the nature of dialogue communication in a mixed modal environment. Finally, we offer concise examples of how current technology may be extended to support the goals above.

Location-Based Services (LBS)

Cellphones have already surpassed the PC as the dominant computing platform (Wright, 2006;Marriner, 2006). In 2006, four in 10 adults browsed the Internet on their wireless handset in Japan, double from 2003 (Wright, 2006). Symbian CEO Nigel Clifford cites the rate of cellphone adoption in India is about 5 million units per month -- PC adoption is only growing at about 5 million a year (Myers, 2006). Reasons cited include both the ease of expanding infrastructure (wireless versus wired) as well as battery performance.

This year, global mobile phone use will top 3.25 billion users (Ridley, Sep 2007) - nearly half the world's population. In Europe and the United States, 4 billion users access maps and download navigation routes (Berg Insight AB [BI], Sep 2007). This is projected to grow to some 43 billion by 2012 (BI, Sep 2007). "The growing adoption will be driven mainly by the intro-

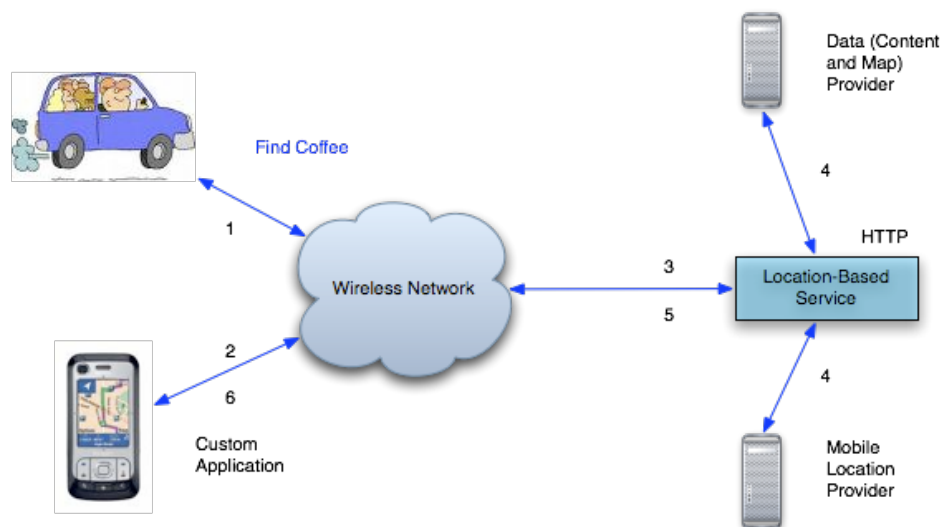
duction of GPS technology in smartphone handsets and bundling of navigation and map content with mobile devices as well as service plans.” (BI, Sep 2007)

Location-Based Services are actually comprised of a set of different technologies. They follow standards promulgated by both the International Standards Organization (ISO) as well as the Open Geospatial Consortium (OGC). The OGC also has an Open Location Services specification which includes five core services: directory service (“spatial yellow page”), gateway service (positioning), location utility service (geocoding), presentation service (map rendering) and routing (OGC, 2005). Depending on the kind of application service offered, LBS subscribe to different content and data providers. For example, data can range from transport timetables, to road network data, to weather data.

There are essentially three methods for a Location-Based Service (LBS) provider to obtain location data from a user’s cellphone: GPS (built-in or wireless), cell tower triangulation, or user reported (Steiniger, Neun & Edwardes, 2006). In the first two cases, LBS providers make agreements with wireless network carriers to receive location data from a cell phone and make it accessible by a web site or call center (Steiniger et al., 2006). Most cellphones today do not allow the user direct access to the GPS data. Location data requires the assistance of the wireless network. A challenge for LBS providers is that they must not only accommodate for different network service providers, but also a wide variety of cellphone operating system and platforms. For this reason, handset users often find that the product they want is not available for their device or network.¹ They are restricted to reporting their position manually.

¹ This paradigm will likely begin to shift soon with formation of the open Handset Alliance (<http://www.openhandsetalliance.com/>).

However, reliance on GPS data also implies that the user must download a custom application built for a specific phone on a specific network -- or be restricted to a specific service offered by the network provider. Thus, end users find they have a limited number of choices depending on phone and network. For any given LBS, only a handful of handsets may be supported and end users are often left in the state of feeling that they wish they could change network provider or cellphone model to take advantage of some particular LBS.



Custom LBS Application

Given these constraints, many cellphone users find themselves in the position of reporting their position manually. Though GPS positioning provides contextual information that enhances the efficiency of location finding, a robust system must also work on older cellphones without GPS or for users that cannot -- or choose not to -- use high speed data transfers. The bottom line is that not all users will have or want a moving map delivered to their cellphone. We must be prepared to deliver location and route information in other ways.

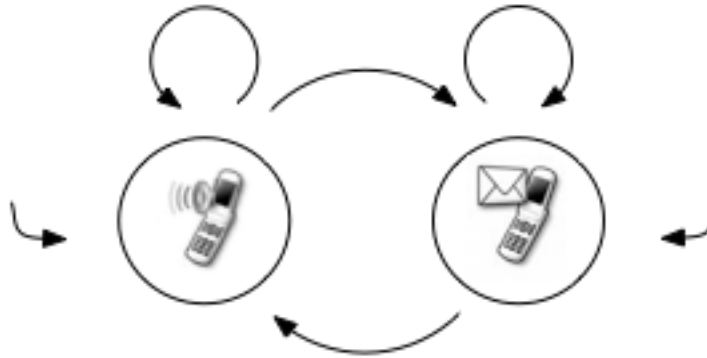
Delivering Interactive Location Assistance on Cellphones

“Based on a June 2006 Harris Interactive study conducted for Tellme, the top three services "on the go" cell phone users are interested in are 1.) maps for driving directions (52%); 2.) voice automated dialing (48%); and 3.) information such as movie listings, stock quotes and weather (33%).” (Tellme Press Release, Aug 21, 2006)

There is strong reason to suggest that LBS, information services, and voice applications will continue on a convergence path and grow in popularity and use. However, compelling interaction does not stop with speech. Imagine any number of painful “conversations” you may have had with menu-driven Interactive Voice Response (IVR) systems. While speech may provide the eyes-free solution many users need, some information (e.g., spatial or memory-intensive) is still better absorbed in a more persistent graphical form. Conversational interaction is not limited to speech.

Many mobile users use text messaging for communication. Links to interactive content can be included in Short Message Service text (SMS). Multimedia Messaging Service (MMS) provides for a combination of media such as text, images and audio. These media may be synchronized for presentation, though there is still yet little interactivity (Openwave, 2002). WAP Push also provides for some degree of rich content but also allows for limited user input and navigation (Openwave, 2002). Not all cellphones support MMS, and WAP Push may be the only option for delivering rich content. However, buyer beware. WAP Push is generally delivered at data rate cost rather than messaging cost. Discovering what the user’s device is capable of, but also factoring in user preferences, is a vital step toward enabling mixed modal communication. Most SMS is delivered in seconds, however, it is possible for delay up to 30 minutes -- though 95% are delivered within 10 seconds (Tanna, 2003; www.text.it, 18/09/03). The user must always feel that he can re-initiate conversation using voice or text, if such a delay occurs.

Given the technologies at hand, conversational exchanges may cross channels as represented in the state diagram below.



Text and Speech Turns

There are at least two well-known commercial systems on the market that provide for fairly fluid voice and SMS information-finding services. Both are large application platforms that give location-based search assistance for businesses. Specifically, the Microsoft subsidiary Tellme (1-800-555-TELL and 83556; <http://www.tellme.com>) and Google (Google Mobile SMS 466453, GOOG-411; http://www.google.com/intl/en_us/mobile/sms/#) span both speech and SMS interaction and provide for very rudimentary conversation. See example transcripts for both Google and Tellme voice services in the Appendix.

Both Tellme and Google voice systems follow the following general model:

get CITY/STATE

get BUSINESS TYPE

get user CHOICE

Tellme also provides other voice service besides search, so prefaces the interaction with a set of user selections including “business search, stock quotes, sports” etc.

Each system has its own strengths and shortcomings. GOOG-411 has exceptional speech recognition quality. The system is very usable -- right up to the end of the interaction where it assumes you will want to be connected to the location you've chosen. This is rather alarming and abrupt, if that is not your goal. To get a text message, the user can interrupt and say “text message” (and then quickly hang-up before a connection is made). System choices are easy to use. Google presents the eight top listings by street name. Users can say the number they'd like to select or use the keyboard for input. When the user selects SMS output, an info text message is sent including a link to a map, which in turn links to driving directions (from the location you earlier specified). The SMS query service does not appear to be integrated with the voice service. Users are not given the option to switch to a voice mode when they interact with the SMS search system.

Tellme.com offers a more integrated service. Speech recognition quality seems lower. However, the main interlocutor's voice quality is quite good and even solicitous when recognition fails. Fortunately, keyed text input allows the user to easily correct errors. Text-to-speech quality for reading of listings is less good. There are multiple voices mixed for any given listing lending a sort of patchwork quality. There are also more prompts with the Tellme system than with Google. However, choice listings are more efficient since if there are multiple chain locations, they are given as: “Starbucks, in multiple cities.” The user can engage in sub-dialogue to hear exact locations for each. Like Google Mobile SMS, Tellme SMS has a very simple dialogue-like functionality such that users can exchange information in small bundles over a sequence of mes-

sages (see SMS transcripts in Appendix). This extends slightly to voice interaction as a user is able to set at least one default preference: send text message automatically.

Modality Choice in Communication

In face-to-face communication, humans naturally communicate through language and action using speech, gaze, facial expression, the movement of lips, head, hands and other kinesthetic channels. Spoken language may be studied and understood in relative isolation, but is complemented by information presented visually.

Current human-computer natural language interfaces are typically limited to unimodal -- or mixed modal² -- text or speech. Empirical studies suggest that natural language interfaces supporting multimodal interaction can be more efficient and natural than those supporting speech-only input (Hauptmann and McAvinney 1993, Oviatt and Olsen 1994, Oviatt 1996, Oviatt and DeAngeli 1998, DeAngeli et al. 1999). Similarly, various media such as text, speech, graphics and animation are best used in combination to convey different types of information in different contexts. For example, instructional manuals generally contain language and graphics to explain how to use technical devices (Wahlster et al. 1993, Feiner and McKeown 1998).

The ability to develop advanced human-computer interfaces that interpret graphical, gestural, and language-based input requires a means for relating information across modalities. One of the problems faced by multimodal systems is the need to resolve cross-modal references. For example, a user may refer to a region on a map display using a natural language expression (“the blue zone”, “the chemical weapons area”, “this box”, etc.) or by pointing and clicking. Such in-

² Mixed modal (also referred to as intermodal) interaction means that one channel or another may be used at any time.

put combines three modalities: the graphical display, the user's gestures, and the user's natural language. A representational system is needed which can access or combine information from across these modalities.

Following Schomaker et al. (1995) and Nigay and Coutez (1993) the term modal covers both the notions of modality as well as mode:

Modality refers to the type of communications channel used to convey or acquire information. It also covers the way an idea is expressed, perceived, or the manner an action is performed. *Mode* refers to a state that determines the way information is interpreted to extract or convey meaning. (Schomaker, 1995)

Frohlich (1991) presents a framework for describing the design space of human-computer interfaces. In addition to the notion of mode and modality, Frohlich refines the notion of mode by identifying media and styles. Examples are in illustrated in table 1 below.

Table 1

Modes by Media and Style

	Language Mode	Action Mode
Media	Speech, text, gesture	sound, graphics, motion
Style	Programming language, command language, natural language, field filling, menu selection	window, iconic, pictorial

Frohlich presents two main advantages to his framework: to clarify terminology used to describe interface design and to classify past and existing systems. Many multimodal researchers give motivations why multiple modes might be more desirable than a single mode (Cohen 1992, Martin 1998, Hauptmann and MacAvinney 1993, Oviatt 1996, Marsh et al. 1994). Individually, each mode has different strengths and weaknesses. Potentially, multiple modes allow users to take

advantage of the strengths of each mode while providing mechanisms for overcoming the weakness of each. The table below illustrates some of these strengths and weakness between modes.

The language mode also allows for semantic information contributed by gesture.

Table 2

Strengths and Weaknesses of Language versus Action

	Language	Action (Direct Manipulation)
Strengths	Intuitive	Intuitive
	Complex queries with the ability to express quantification, attribute and object relations, negation, counterfactuals, categorization, ordering, and aggregate operations	Easy to manipulate spatial properties of objects (size, shape, placement)
	Discourse context permits mechanisms such as anaphora, ellipsis, and deixis	Immediate feedback
		Easy to remember how to operate on objects in spatial metaphor; not ambiguous or error prone
Weaknesses	Not appropriate for all tasks	Not appropriate for all tasks
	What can you say and how can you say it?	Complex tasks including batch operations or operations on multiple objects may require many actions: previous operations cannot be referred to
	Error prone	
	Ambiguous at times	Can't express relations well

The same researchers provide reasons why users might prefer to use a combination of speech and gesture for different tasks: naturalness, economy, and clarity. It seems obvious that flexible modality choice for both system input and output might be more advantageous than sin-

gle mode operation. But why users choose to utter one expression over another is a complex question. Language use in a multimodal environment is at least a function of the nature of the task, complexity of the task, user experience, user preference and affordances of the system. We suggest that the ideal communication with a cellphone device is *fleximodal*³ where users are able to rapidly and fluidly change modalities depending on the context at hand.

All standard computer interfaces today are at least mixed modal. In a typical graphical environment a user communicates through a visual medium using keyboard textual input and mouse gestures. Most applications support user input from a variety of methods. We are specifically concerned with language input from across two channels. Referents in one channel may be referred to in the other. And users may move fluidly between channels depending on the task and relative cognitive (or physical) effort to deliver or respond to information in a given channel.

Conversational Dialogue Systems

According to Clark and Schaeffer (1989), grounding is the process of establishing and maintaining common ground (mutual belief) in conversation. **Grounding** is the means by which “new” information gets added to the common ground; participants try to establish that what they’ve said is what has been understood. Furthermore, grounding is a joint process in which conversational participants collaborate on understanding. To ground information is to make it part of the common ground. Implicit to a theory of grounding is the notion that understanding can never be perfect. Clark’s theory of grounding is intended not only to account for the *process* of adding information to the common ground, but also to account for what counts as a miscommunication and what counts as a repair.

³ This term is coined from a paper by Meyers et al. (2002)

Clark's work is founded on a long tradition of conversational analysis (Sacks, Schegloff, Jefferson, etc.) as well as theories of mutual belief (Lewis 1969; Schiffer 1972) and common ground (Stalnaker 1979; Clark and Marshall 1981). Clark and Schaefer (1987; 1989) proposed a formal model characterized by three key elements:

- Conversation proceeds at two levels: one contains *topical content*, one contains *grounding of content*;
- Contributions to conversation are organized hierarchically and may be embedded;
- People reach satisfaction in dialogue according to specification of the *grounding criterion* ("the speaker and addressees mutually believe that the addressees have understood what the speaker meant to a criterion sufficient for current purposes").

The most basic element in conversation is the *contribution*. Each contribution has two phases – the **presentation phase** (where a speaker A presents an utterance) followed by an **acceptance phase** (where an addressee B presents evidence of understanding). Because we're talking about mutual acceptance, in B's acceptance, B should let know A what state he's in and for which part of the utterance. Also, in the acceptance phase, A implicitly (by allowing the conversation to proceed) or explicitly accepts B's acceptance.

Several researchers have proposed extensions to Clark's theory of grounding for use in human-computer interaction (Cahn 1992; Traum and Hinkelman 1992; Traum 1994; Traum and Allen 1994; Brennan and Hulstien 1995; Cahn and Brennan 1999). According to these researchers, there are several practical advantages to using Clark's formal model:

- Tracking understanding in dialogue;
- Handling miscommunications;

- Use of an explanatory model that also provides a representation of the *process* of dialogue.

Research conversational systems employ a system component known as a dialogue manager. The dialogue manager performs the role of understanding but also deciding on an appropriate response. Dialogue managers model the process of human-human dialogue. Because any human communication presumes miscommunication, a dialogue manager must also be concerned with avoiding miscommunications as well as detecting and repairing them. There are a variety of different sorts of dialogue theories depending on the kind of dialogue modeled (e.g., tutorial, negotiative, information-seeking, action-oriented, etc.). There are also a variety of computational models that dialogue systems employ (e.g., finite-state, plan-based, agent-based and frame-based). Various computational and theoretical theories are proposed to account for variable features and complexities associated with different types of dialogues.

Ginzburg (1997) poses an issue-based model for grounding called Questions Under Discussion (QUD). QUD is concerned with grounding information as meaning questions (“what do you mean”) and acceptance questions (“should *u* be accepted?”). A proposition can be accepted as either a fact or as a topic (issue) of discussion. A question is always a topic of discussion. A dialogue manager modeling QUD is ideal for question answering tasks such as ours. Question accommodation allows the system to understand answers addressing issues which have not yet been raised. In cases of ambiguity, clarification dialogues may be needed. Larsson’s (2002) issue-based dialogue manager is well suited to answering questions such as “where can I find coffee?”. Like other sophisticated dialogue managers, it is designed to also account for common

dialogue phenomena such as interactive grounding (verification), accommodation⁴, tracking multiple conversational threads and mixed initiative (Larsson 2007).

Larsson (2007) reports that although some basic mechanisms for dealing with unrequested information exists in VoiceXML 2.1, there is currently no support for clarification sub-dialogues⁵. Also, VoiceXML 2.1 provides only very limited support for information sharing between subtasks and dealing with several simultaneous tasks. VoiceXML has practical limitations, as well. It mixes dialogue knowledge, domain knowledge and language knowledge into a single specification. It is a finite-state approach to dialogue and can be brittle and difficult to debug. It also does not yet support multimodal interaction. However, VoiceXML 3.0 intends to correct some of these deficiencies (Larsson 2007). Larsson is in the process of transforming his dialogue architecture to a VoiceXML 3.0 State Chart XML.

Conversational Interaction on Cellphones

Unlike Interactive Voice Response (IVR) systems which model menu-based dialogue, conversational systems are generally concerned with modeling natural dialogue interaction⁶. Useful features of dialogue interaction over menu-based interaction are (Larsson, Amores, Karagjosova, Milward, Tsovaltzi, 2002):

- The user and system can provide information in any order

⁴ In semantic accommodation, if the user adds information that is not explicitly requested, the system accommodates those propositions into the existing plan. Accommodation is described by Lewis (1979).

⁵ Dialogue designers can creatively avoid clarification by choosing alternative strategies such as verification.

⁶ The DICO project has explored the use of menu-based interaction with a dialogue manager for the control of an MP3 (Larsson and Villing, 2007; Villing, 2007, Villing and Larsson, 2006).

- The task does not have to be pre-specified
- The user does not have to learn menu structure
- The user can switch tasks and engage in multiple tasks simultaneously

Potential disadvantages are:

- The user may feel confused about what to say if not directed at all points in the interaction
- The user also may not cognitively share the same task model as the system

Typically, IVR systems are able to process language to some degree, though they may have no representation of user task (Hocek, 2002). Task-based dialogue, on the other hand, is concerned with understanding the user's task as well as the business of managing turns (Larsson 2002)⁷. For our purposes, we are concerned with the use of VoiceXML for speech interaction since it's an industry standard that supports plan constructs and allows for multiple active grammars. Our information-seeking dialogue is best modeled as a form-filling dialogue in VoiceXML. However, text interaction provides for subtle extensions that improve user experience, yet maintaining a unified and consistent interaction experience.

SMS text messaging has many attributes well suited to dialogue interaction. Participants in SMS exchanges engage in turn taking. Though messages are restricted to 160 characters, they are also designed to link to one another in batches so that they may vary in form, content and length (Openwave, 2002). Typically, text messages are delivered within seconds, though on occasion a message may be delayed. However, unlike spoken communication, users engaged in messaging don't expect their message to be accepted immediately. It is not a completely synchronous form of communication. Notwithstanding, users engage in turn-taking behavior just as they do in oral

⁷ Information-seeking is a sort of task-based dialogue.

communication. But timing and interruption are less relevant. And it is also wholly possible for a user to ignore a message and thus not ground its information content. However, the sender can follow up with additional attempts if no response occurs during some expected period of time.

SMS texting has some advantage over voice interaction for location finding. Messages are persistent and that means directions can be stored and retrieved during the course of navigation. MMS is an advancement over SMS that also supports the transmission of spatial information (e.g., maps) and synchronized media presentation.

Text messaging also affords a user a great deal of flexibility in constructing a message. For example,

User: *cumberland maryland sort popular tags fireplace*

In this message the user not only requests information for coffee shops in a specific location, but also adds information about how the system should present that information. This is a simple example of task and question accommodation. One task is to find coffee shops and the other is to present them. Information about how to present them is question accommodation.

Another characteristic of text messaging that reflects characteristics of spoken language is elliptical pronominal references. For example,

System: *Respond with a number and any of the following: directions, details, reviews, map, tags*

User: *tags 1*

System: *Coffee Hut in Rocky Gap. Tags: fireplace, soymilk*

User: *directions*

As in spoken communication, the referent of “directions” is the entity most in focus and in an accessible location in dialogue structure. Unlike speech communication, large time gaps may not affect the cognitive accessibility of the referent (though I found no evidence in the literature to either support or dispute this). However, it is possible that ambiguity may occur, given that message transmission cannot be guaranteed to be strictly serial.

Note, finally, that the ability to add information using task and question accommodation can dramatically improve the relevance of search results. Both Tellme and GOOG-411 assume that the primary factor for a user’s decision to select a particular choice is location. However, in our exploratory design study we found that our users were less than satisfied with the results they got from commercial services. They expressed a desire to have search returns filtered on other criteria than location alone. Discussion of our social tagging and recommender features go well beyond the scope of this literature review. However, the conversational nature of our interaction supports concepts such as progressive (incremental) filtering and flexible accommodation as in the examples above.

Extending State-of-the-Art Systems

Despite limitations of VoiceXML, text messaging formats (since MMS is not yet universal) and the availability of LBS, it is possible to make fairly significant improvements to both the efficiency of location search as well as the relevance of search returns. Below are a number of specific observations made by closely examining Tellme and Google voice and SMS location search.

- Avoid presuming that a user's search criteria is based solely on location. Both Tellme and GOOG-411 voice systems appear to make this presumption. Users may need review results in spatial or social contexts before requesting directions to one or another.
- Extend conversational interaction by using dialogue context such that users can incrementally filter search results and pursue multiple search alternatives without losing the context of the broader search. There is a hint of this in both the SMS systems provided by Tellme and Google. However, once a user selects a particular location, the query is considered complete. There is no way for the user to back up to his original goal. Given limitations of VoiceXML, it is possible this may prove to be serious technical challenge. This should, however, be quite achievable in SMS interaction
- Allow for semantic accommodation and thus enable a highly efficient text search capability. This should include resolving pronominal reference as well as allowing for the user to add additional information that provides answers to yet unanswered questions.
- Make few assumptions about whether a user should use text or voice at any point during an interaction. The interaction should feel seamless and the user choose the mode depending on his cognitive state and the task at hand. The user should not feel as if interaction styles or task plans have changed.
- Improve the relevance of search by incorporating user-contributed data (e.g., ratings, comments, photos) and allowing users to search and sort along these criteria.

There are a number of direction we imagine will be useful to integrate with additional research.

- Use GPS location data dynamically during conversational interaction, instead of passing a user off to a separate GPS mapping service. Such information can be used in a variety of ways. For example, to recognize when a user has reached a destination. For now, naive users are averse to installing custom software and this is not necessarily the best option.
- There is much potential in increasing social services by enabling seamless sharing of location with friends by auto SMS updates and a shared map.
- Provide for an online chat recommender system. If a user is in a new city, he may wish to consult with locals. It may be useful to include a web chat capability to speak with other users more directly. It may also be useful to mine web chat conversation for messages pertaining to a particular query response.
- Work with other vendors and standards bodies to address limitations of VoiceXML and SMS technologies.
- Encourage the development and use of a threaded conversation management tool for both SMS and voice messages. Currently, cellphone providers see these two modalities as separate and it's not possible to integrate the two in a single log.
- Finally, though this was not discussed above, to ensure that the interaction style and strategies are consistent across modalities, generating both the VoiceXML and SMS dialogue recipes from a single source may be valuable to ensuring user experience, dialogue design, and capabilities development are locked in step.

Looking toward future research in this arena, there is much to learn about how technology affordances affect personalization, dialogue behavior, and modality preference. As

technologies advance, we must continually assess system strategies and techniques for communicating with the user.

Appendix

Goog-411 Transcript

Prompt: What city and state

User: Clarksburg, Maryland

Prompt: What business name or category?

User: coffee

Prompt: Searching

Prompt:

Top eight results

Number 1 Myorga Coffee on Stringtown Road

To select number one, you can press one or say number one

Number 2 Starbucks on Frederick Road, Germantown

Number 3 Starbucks on Frederick Road, Germantown

Number 4 Starbucks Coffee on Germantown Road, Germantown

To start a new search, say start over anytime

Number 5 Starbucks on Frederick Road, Germantown

Number 6 Dunkin Donuts on Wisteria Drive, Germantown

Number 7 Mayorga Coffee on Wisteria Drive, Germantown

Number 8, Aqui Brazilian Coffee on Wisteria Drive, Germantown

Top of the list

User: <Punch 8>

Number 1 -

Prompt: Number 8, Aqui Brazilian Coffee on Wisteria Drive, Germantown

Prompt: I'll connect you or say details or go back

User: Text message (able to do this by reading instructions on website)

Prompt: Hold on, Sending text message

Prompt: Hang on and I'll connect you, or say go back.

Map output and driving directions from Google:

Google Local BETA

Zoom: ± | 3100 ft | Change

Start from **Clarksburg, MD**
 Arrive at **12615 Wisteria Dr, Germantown, MD 20874**
 Distance **6.1 mi**

1. Head **northwest** on **Frederick Rd/MD-355** toward **Clarksburg Rd/MD-121 - 246 ft**
2. Turn **left** at **Clarksburg Rd/MD-121 - 0.6 mi**
3. Merge onto **I-270 S** via the ramp to **Washington - 3.6 mi**
4. Take exit **15B-A** to merge onto **Germantown Rd/MD-118 S** toward **Germantown - 1.0 mi**
5. Turn **left** at **Middlebrook Rd - 0.4 mi**
6. Turn **right** at **Great Seneca Hwy/MD-119 S - 0.3 mi**
7. Turn **left** at **Wisteria Dr - 0.1 mi**

[Driving directions](#)

Business details

Aqui Brazilian Coffee
 120 mi E
 12615 Wisteria Dr, Germantown, MD
 (301) 540-3395
[« Back to results](#)

Aqui Brazilian Coffe

lisah2u@gmail.com
[Google Home](#)
[Sign out](#)
[Settings](#)
[Leave us feedback](#)
[Help & Terms](#)

View Google in:
[Mobile](#) | [Classic](#)
 ©2007 Google

What

Where

[Driving Directions](#)

[Google Home](#)
[Help & Terms](#)
 ©2007 Google

Tellme Voice Transcript

Prompt: Main menu

Prompt: business search, ring tones, stock quotes, news center, sports, entertainment, travel, and popular shortcuts

User: business search

Prompt: hmmm. I didn't get that. Please say one of our categories or say main menu

Prompt: business search

Prompt: I'm sorry, I still didn't get that. To hear all the categories we have, say main menu

User: business search

Prompt: I'm sorry. I'm having trouble understanding. You can also punch the first three letters of one of our categories on the telephone keypad, for instance -- okay, business search then.

User: <punch BUS>

Prompt: city and state please

User: clarksburg, maryland

Prompt: okay. what business, or type of business are you looking for?

User: coffee

Prompt: To narrow your search, you can say by intersection or you can say I don't know

Prompt: Alright. When you hear the one you want, just say the number.

Prompt:

1) Mayorga coffee

2) Starbucks, in multiple cities - (There are two. Do you want the ___ or the ___)

3) Music Cafe, in Damascus

Prompt: You can say repeat, tell me more choices, or new search

Prompt: Sorry I didn't hear you.

User: Tell me more choices.

Prompt: More choices

Prompt:

4) Dunkin Donuts in multiple cities

5) HUUI Brazilian Coffee, in Germantown

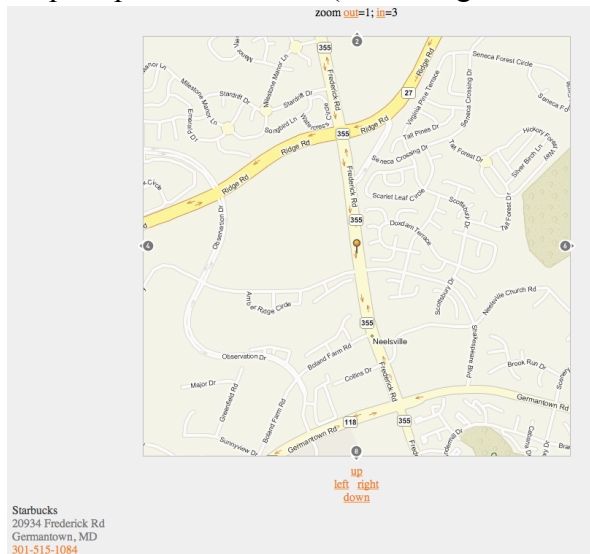
User: 5

Prompt: Okay, HUUI Brazilian Coffee at the 12615 Wisteria Drive Germantown Square

Prompt: The number is 301-540-3395. I'll text you that info now.

Prompt: You can say share this listing, you can also say repeat the info or start over.

Map output from Tellme (no driving directions):



References

- Angeli, A. D., F. Wolff, et al. (1999). Relevance and Perceptual Constraints in Multimodal Referring Actions. Proceedings of the Workshop on Deixis, Demonstration and Deictic Belief at ESSLLI X.
- B.A. Myers et al. (2002). "Fleximodal and Multimachine User Interfaces", Proc. IEEE 4th Int'l Conf. Multimodal Interfaces, IEEE Press, 2002, pp. 343-348.
- Berg Insight AB, (Sep 2007). Mobile Maps and Navigation. Retrieved from:
http://www.researchandmarkets.com/reportinfo.asp?report_id=553108&t=e&cat_id=
- Brennan, S. E. and E. A. Hulteen (1995). "Interaction and feedback in a spoken language system: A theoretical framework." Knowledge-Based Systems 8: 143-151.
- Cahn, J. (1992). A Computational Architecture for Mutual Understanding in Dialog, Music and Cognition Group, M.I.T. Media Laboratory.
- Cahn, J. E. and S. E. Brennan (1999). A Psychological Model of Grounding and Repair in Dialog. Proceedings of the Fall 1999 AAAI Symposium on Psychological Models of Communication in Collaborative Systems, Sea Cliff, Massachusetts.
- Clark, H. H. and C. R. Marshall (1981). Definite Reference and Mutual Knowledge. Elements of Discourse Understanding. A. K. Joshi, B. L. Webber and I. A. Sag. New York, Cambridge University Press.
- Clark, H. H. and E. F. Schaefer (1987). "Collaborating on Contributions to Conversations." Language and Cognitive Processes 2(1): 19-41.
- Clark, H. H. and E. F. Schaefer (1989). "Contributing to Discourse." Cognitive Science 13: 259-284.
- Cohen, P. R. (1992). The Role of Natural Language in a Multimodal Interface. UIST '92, Proceedings of the ACM Symposium on User Interface Software and Technology, Monterey, CA.
- Feiner, S. K. and K. R. McKeown (1991). "Automating the Generation of Coordinated Multimedia Explanations." IEEE Computer 24(10): 33-41.
- Frohlich, D. M. (1991). The Design Space of Interfaces. Multimedia: Systems, Interaction and Applications, Proceedings of the First Eurographics Workshop, Stockholm.
- Hauptmann, A. G. and P. McAvinney (1993). Gestures with Speech for Graphics Manipulation. Intl. J. Man-Machine Studies 38: 231-249.
- Hocek, A. (2002). "VoiceXML and Next-Generation Voice Services". XML 02 Conference. Baltimore, MD. Retrieved from:
http://www.idealliance.org/papers/xml02/dx_xml02/html/abstract/06-02-01.html
<http://www.text.it/mediacentre/default.asp?intPageID=132> [Accessed 18/09/03]
- Jessica Villing (2007): DICO: Drive and Talk. Proceedings of the HCI International 2007 (HCII, Beijing, China 22-27 July 2007).
- Jessica Villing, Staffan Larsson (2006). Dico: a Multimodal Menu-based In-vehicle Dialogue System In Sclangen and Fernández (Eds.): brandial'06 Proceedings of the 10th Workshop on the Semantics and Pragmatics of Dialogue (Sem-Dial 10, Potsdam, Germany).
- Jonathan Ginzburg. (May 1997). "Querying and Assertions in Dialogue." Draft paper.

- Larsson, S., Amores, G., Karagjosova, E., Milward, D., Tsovaltzi, D. (2002). "Flexible Dialogue." Technical Report SIRidUS project deliverable D1.4.
- Lewis, D. (1979). "Scorekeeping in a Language Game." Journal of Philosophical Logic 8: 339-359.
- Lewis, D. A. (1969). Convention: A Philosophical Study, Harvard University Press.
- Marriner, L. (2006, Oct 18). CNET News, Cell phone already surpasses PC as dominant computing platform: reader comment, Retrieved from: <http://www.news.com/5208-1039-0.html?forumID=1&threadID=22079&messageID=194463&start=-1&reply=true>
- Marsh, E., K. Wauchope, et al. (1994). Human-Machine Dialogue for Multi-Modal Decision Support Systems.
- Martin, J.-C., L. Julia, et al. (1998). A Theoretical Framework for Multimodal User Studies. CMC 98, Tilberg, The Netherlands.
- Myers, D. (2006, Oct 17). CNET News, Symbian forecasts the death of the PC, Oct 17, 2006. Retrieved from: <http://www.news.com/2100-1039-6126565.html?tag=tb>
- Nigay, L. and J. Coutaz (1993). A Design Space for Multimodal Systems - Concurrent Processing and Data Fusion. INTERCHI '93 - Conference on Human Factors in Computing Systems, Amsterdam, Addison, Wesley.
- Open Geospatial Consortium, (2005). Open Location Services 1.1.
- Openwave. (2002). Comparison of WAP Push and Short Message Service. Retrieved from: http://developer.openwave.com/docs/wappush_vs_sms.pdf
- Oviatt, S. L. (1996). Multimodal Interfaces for Dynamic Interactive Maps. Proceedings of Conference on Human Factors in Computing Systems: CHI '96, ACM Press.
- Oviatt, S. L. and E. Olsen (1994). Integration Themes in Multimodal Human-Computer Interaction. Proceedings of the International Conference on Spoken Language Processing.
- Oviatt, S. L., A. DeAngeli, et al. (1998). Integration and Synchronization of Input Modes during Multimodal Human-Computer Interaction. Proceedings of Conference on Human Factors in Computing Systems: CHI '97, ACM Press.
- Ridley, K. (2007, Jun 27). Global Mobile Phone Use to Hit Record 3.25 Billion. Retrieved from: <http://www.reuters.com/article/companyNewsAndPR/idUSL2712199720070627>
- Schiffer, S. R. (1972). Meaning, Oxford University Press.
- Schomaker, L., J. Nijtmans, et al. (1995). A Taxonomy of Multimodal Interaction in the Human Information Processing System.
- Staffan Larsson (2002). Issue-based Dialogue Management. PhD Thesis, Goteborg University.
- Staffan Larsson (2007): Rapid Prototyping using Issue-based Dialogue Management in GoDiS. Workshop on Advanced Dialogs. VoceXML Forum Tools Committee, Advanced Dialogs Working Group.
- Staffan Larsson, Jessica Villing (2007). The DICO project: A Multimodal Menu-based In-vehicle Dialogue System. In Bunt, H.C., and Thijsse, E. C. G. (eds): Proceedings of the 7th International Workshop on Computational Semantics (IWCS-7, Tilburg, The Netherlands).
- Stalnaker (1979). Assertion. Syntax and Semantics. P. Cole, Academic Press. 9: 315-332.

- Steiniger, S., Neun, N., Edwardes, A. (2006). Foundations of Location Based Services, Lesson 1, Lecture Notes on LBS, V. 1.0. Retrieved from:
http://www.geo.unizh.ch/publications/cartouche/lbs_lecturenotes_steinigeretal2006.pdf
- Tanna, V. (2004). The Turn-taking System of Short Message Service (SMS) Exchange. Dissertation.
- Tellme. (2006, Aug 21). Tellme Surpasses 100 Million Voice Searches Per Month. Retrieved from: <http://www.tellme.com/about/PressRoom/release/20060821>
- Traum, D. R. (1994). A Computational Theory of Grounding in Natural Language Conversation, University of Rochester.
- Traum, D. R. and E. Hinkelman (1992). "Conversation Acts in Task-Oriented Spoken Dialogue." *Computational Intelligence* 8(3): 575-599.
- Traum, D. R. and J. F. Allen (1994). A Speech Acts Approach to Grounding in Conversation. Proceedings 2nd International Conference on Spoken Language Processing (ICSLP-92).
- Wahlster, W., E. Andre, et al. (1993). "Plan-Based Integration of Natural Language and Graphics Generation." *Artificial Intelligence* 63(1-2): 387-427.
- Wright, A. (2006, Apr 18). Mobile Phones Could Soon Rival the PC as the World's Dominant Internet Platform. Retrieved from:
<http://www.ipsos-na.com/news/pressrelease.cfm?id=3049>.
www.text.it 'Fast Text Facts'.