Multimodal Applications from Mobile to Kiosk

Michael Johnston
AT&T Research
180 Park Avenue
Florham Park, NJ 07932
johnston@research.att.com

Srinivas Bangalore
AT&T Research
180 Park Avenue
Florham Park, NJ 07932
srini@research.att.com

Abstract

Multimodal interfaces provide more flexible and compelling interaction and can enable public information kiosks to support more complex tasks for a broader community of users. MATCHKiosk is a multimodal interactive city guide which provides users with the freedom to interact using speech, pen, touch or multimodal inputs. The system responds by generating multimodal presentations that synchronize synthetic speech with a life-like virtual agent and dynamically generated graphics.

1 Introduction

Since the introduction of automated teller machines in the late 1970s, public kiosks have been introduced to provide users with automated access to a broad range of information, assistance, and services. These include self check-in at airports, ticket machines in railway and bus stations, directions and maps in car rental offices, interactive tourist and visitor guides in tourist offices and museums, and more recently, automated check-out in retail stores. The majority of these systems provide a rigid structured graphical interface and user input by only touch or keypad, and as a result can only support a small number of simple tasks. As automated kiosks become more commonplace and have to support more complex tasks for a broader community of users, they will need to provide a more flexible and compelling user interface.

One major motivation for developing multimodal interfaces for mobile devices is the lack of a keyboard or mouse (Oviatt and Cohen, 2000; Johnston and Bangalore, 2000). This limitation is also true of many different kinds of public information kiosks where security, hygiene, or space concerns make a physical keyboard or mouse impractical. Also, mobile users interacting with kiosks are often encumbered with briefcases, phones, or other equipment, leaving only one hand free for interaction. Kiosks often provide a touchscreen for input, opening up the possibility of an onscreen keyboard, but these can be awkward to use and occupy a considerable amount of screen real estate, generally leading to a more moded and cumbersome graphical interface.

A number of experimental systems have investigated adding speech input to interactive graphical kiosks (Raisamo, 1998; Gustafson et al., 1999; Narayanan et al., 2000; Lamel et al., 2002). Other work has investigated adding both speech and gesture input (using computer vision) in an interactive kiosk (Wahlster, 2003; Cassell et al., 2002).

We describe MATCHKiosk, (Multimodal Access To City Help Kiosk) an interactive public information kiosk with a multimodal interface which provides users with the flexibility to provide input using speech, handwriting, touch, or composite multimodal commands combining multiple different modes. The system responds to the user by generating multimodal presentations which combine spoken output, a life-like graphical talking head, and dynamic graphical displays. MATCHKiosk provides an interactive city guide for New York and Washington D.C., including information about restaurants and directions on the subway or metro. It develops on our previous work on a multimodal city guide on a mobile tablet (MATCH) (Johnston et al., 2001; Johnston et al., 2002b; Johnston et al., 2002a). The system has been deployed for testing and data collection in an AT&T facility in Washington, D.C. where it provides visitors with information about places to eat, points of interest, and getting around on the DC Metro.

2 The MATCHKiosk

The MATCHKiosk runs on a Windows PC mounted in a rugged cabinet (Figure 1). It has a touch screen which supports both touch and pen input, and also contains a printer, whose output emerges from a slot below the screen. The cabinet also contains speakers and an array microphone is mounted above the screen. There are three main components to the graphical user interface (Figure 2). On the right, there is a panel with a dynamic map display, a click-to-speak button, and a window for feedback on speech recognition. As the user interacts with the system the map display dynamically pans and zooms and the locations of restaurants and other points of interest, graphical callouts with information, and subway route segments are displayed. In

Figure 1: Kiosk Hardware

the top left there is a photo-realistic virtual agent (Cosatto and Graf, 2000), synthesized by concatenating and blending image samples. Below the agent, there is a panel with large buttons which enable easy access to help and common functions. The buttons presented are context sensitive and change over the course of interaction.

Figure 2: Kiosk Interface

The basic functions of the system are to enable users to locate restaurants and other points of interest based on attributes such as price, location, and food type, to request information about them such as phone numbers, addresses, and reviews, and to provide directions on the subway or metro between locations. There are also commands for panning and zooming the map. The system provides users with a high degree of flexibility in the inputs they use in accessing these functions. For example, when looking for restaurants the user can employ speech e.g. *find me moderately priced italian restaurants in Alexandria*, a multimodal combination of speech and pen, e.g. *moderate italian restaurants in this area* and circling Alexandria on the map, or solely pen, e.g. user writes *moderate italian* and *alexandria*. Similarly, when requesting directions they can use speech, e.g. *How do I get to the Smithsonian?*, multimodal, e.g. *How do I get from here to here?* and circling or touching two locations on the map, or pen, e.g. in Figure 2 the user has circled a location on the map and handwritten the word *route*.

System output consists of coordinated presentations combining synthetic speech with graphical actions on the map. For example, when showing a subway route, as the virtual agent speaks each instruction in turn, the map display zooms and shows the corresponding route segment graphically. The kiosk system also has a print capability. When a route has been presented, one of the context sensitive buttons changes to **Print Directions**. When this is pressed the system generates an XHTML document containing a map with step by step textual directions and this is sent to the printer using an XHTML-print capability.

If the system has low confidence in a user input, based on the ASR or pen recognition score, it requests confirmation from the user. The user can confirm using speech, pen, or by touching on a checkmark or cross mark which appear in the bottom right of the screen. Context-sensitive graphical widgets are also used for resolving ambiguity and vagueness in the user inputs. For example, if the user asks for the Smithsonian Museum a small menu appears in the bottom right of the map enabling them to select between the different museum sites. If the user asks to see restaurants near a particular location, e.g. *show restaurants near the white house*, a graphical slider appears enabling the user to fine tune just how near.

The system also features a context-sensitive multimodal help mechanism (Hastie et al., 2002) which provides assistance to users in the context of their current task, without redirecting them to separate help system. The help system is triggered by spoken or written requests for help, by touching the help buttons on the left, or when the user has made several unsuccessful inputs. The type of help is chosen based on the current dialog state and the state of the visual interface. If more than one type of help is applicable a graphical menu appears. Help messages consist of multimodal presentations combining spoken output with ink drawn on the display by the system. For example, if the user has just requested to see restaurants and they are now clearly visible on the display, the system will provide help on getting information about them.

3 Multimodal Kiosk Architecture

The underlying architecture of MATCHKiosk consists of a series of re-usable components which

communicate using XML messages sent over sockets through a facilitator (MCUBE) (Figure 3). Users interact with the system through the Multimodal UI displayed on the touchscreen. Their speech and ink are processed by speech recognition (ASR) and handwriting/gesture recognition (GESTURE, HW RECO) components respectively. These recognition processes result in lattices of potential words and gestures/handwriting. These are then combined and assigned a meaning representation using a multimodal language processing architecture based on finite-state techniques (MMFST) (Johnston and Bangalore, 2000; Johnston et al., 2002b). This provides as output a lattice encoding all of the potential meaning representations assigned to the user inputs. This lattice is flattened to an N-best list and passed to a multimodal dialog manager (MDM) (Johnston et al., 2002b) which re-ranks them in accordance with the current dialogue state. If additional information or confirmation is required, the MDM uses the virtual agent to enter into a short information gathering dialogue with the user. Once a command or query is complete, it is passed to the multimodal generation component (MMGEN), which builds a multimodal *score* indicating a coordinated sequence of graphical actions and TTS prompts. This score is passed back to the Multimodal UI. The Multimodal UI passes prompts to a visual text-to-speech component (Cosatto and Graf, 2000) which communicates with the AT&T Natural Voices TTS engine (Beutnagel et al., 1999) in order to coordinate the lip movements of the virtual agent with synthetic speech output. As prompts are realized the Multimodal UI receives notifications and presents coordinated graphical actions. The subway route server is an application server which identifies the best route between any two locations.

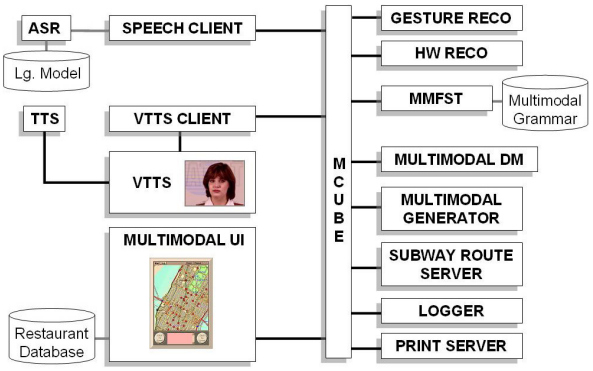

Figure 3: Multimodal Kiosk Architecture

4 Discussion and Related Work

A number of design issues arose in the development of the kiosk, many of which highlight differences between multimodal interfaces for kiosks and those for mobile systems.

Array Microphone While on a mobile device a close-talking headset or on-device microphone can be used, we found that a single microphone had very poor performance on the kiosk. Users stand in different positions with respect to the display and there may be more than one person standing in front. To overcome this problem we mounted an array microphone above the touchscreen which tracks the location of the talker.

Robust Recognition and Understanding is particularly important for kiosks since they have so many first-time users. We utilize the techniques for robust language modelling and multimodal understanding described in Bangalore and Johnston (2004).

Social Interaction For mobile multimodal interfaces, even those with graphical embodiment, we found there to be little or no need to support social greetings and small talk. However, for a public kiosk which different unknown users will approach those capabilities are important. We added basic support for social interaction to the language understanding and dialog components. The system is able to respond to inputs such as *Hello*, *How are you?*, *Would you like to join us for lunch?* and so on.

Context-sensitive GUI Compared to mobile systems, on palmtops, phones, and tablets, kiosks can offer more screen real estate for graphical interaction. This allowed for large easy to read buttons for accessing help and other functions. The system alters these as the dialog progresses. These buttons enable the system to support a kind of mixed-initiative in multimodal interaction where the user can take initiative in the spoken and handwritten modes while the system is also able to provide a more system-oriented initiative in the graphical mode.

Printing Kiosks can make use of printed output as a modality. One of the issues that arises is that it is frequently the case that printed outputs such as directions should take a very different style and format from onscreen presentations.

In previous work, a number of different multimodal kiosk systems supporting different sets of input and output modalities have been developed. The Touch-N-Speak kiosk (Raisamo, 1998) combines spoken language input with a touchscreen. The August system (Gustafson et al., 1999) is a multimodal dialog system mounted in a public kiosk. It supported spoken input from users and multimodal output with a talking head, text to speech, and two graphical displays. The system was deployed in a cultural center in Stockholm, enabling collection of realistic data from the general public. SmartKom-Public (Wahlster, 2003) is an interactive public information kiosk that supports multimodal input through speech, hand gestures, and facial expressions. The system uses a number of cameras

and a video projector for the display. The MASK kiosk (Lamel et al., 2002) , developed by LIMSI and the French national railway (SNCF), provides rail tickets and information using a speech and touch interface. The mVPQ kiosk system (Narayanan et al., 2000) provides access to corporate directory information and call completion. Users can provide input by either speech or touching options presented on a graphical display. MACK, the Media Lab Autonomous Conversational Kiosk, (Cassell et al., 2002) provides information about groups and individuals at the MIT Media Lab. Users interact using speech and gestures on a paper map that sits between the user and an embodied agent.

In contrast to August and mVPQ, MATCHKiosk supports composite multimodal input combining speech with pen drawings and touch. The SmartKom-Public kiosk supports composite input, but differs in that it uses free hand gesture for pointing while MATCH utilizes pen input and touch. August, SmartKom-Public, and MATCHKiosk all employ graphical embodiments. SmartKom uses an animated character, August a model-based talking head, and MATCHKiosk a sample-based video-realistic talking head. MACK uses articulated graphical embodiment with ability to gesture. In Touch-N-Speak a number of different techniques using time and pressure are examined for enabling selection of areas on a map using touch input. In MATCHKiosk, this issue does not arise since areas can be selected precisely by drawing with the pen.

5 Conclusion

We have presented a multimodal public information kiosk, MATCHKiosk, which supports complex unstructured tasks such as browsing for restaurants and subway directions. Users have the flexibility to interact using speech, pen/touch, or multimodal inputs. The system responds with multimodal presentations which coordinate synthetic speech, a virtual agent, graphical displays, and system use of electronic ink.

Acknowledgements Thanks to Eric Cosatto, Hans Peter Graf, and Joern Ostermann for their help with integrating the talking head. Thanks also to Patrick Ehlen, Amanda Stent, Helen Hastie, Guna Vasireddy, Mazin Rahim, Candy Kamm, Marilyn Walker, Steve Whittaker, and Preetam Maloor for their contributions to the MATCH project. Thanks to Paul Burke for his assistance with XHTML-print.

References

S. Bangalore and M. Johnston. 2004. Balancing Data-driven and Rule-based Approaches in the Context of a Multimodal Conversational System. In *Proceedings of HLT-NAACL*, Boston, MA.

M. Beutnagel, A. Conkie, J. Schroeter, Y. Stylianou, and A. Syrdal. 1999. The AT&T Next-Generation TTS. In *In Joint Meeting of ASA; EAA and DAGA*.

J. Cassell, T. Stocky, T. Bickmore, Y. Gao, Y. Nakano, K. Ryokai, D. Tversky, C. Vaucelle, and H. Vilhjalmsson. 2002. MACK: Media lab autonomous conversational kiosk. In *Proceedings of IMAGINA02, Monte Carlo*.

E. Cosatto and H. P. Graf. 2000. Photo-realistic Talking-heads from Image Samples. *IEEE Transactions on Multimedia*, 2(3):152–163.

J. Gustafson, N. Lindberg, and M. Lundeberg. 1999. The August spoken dialogue system. In *Proceedings of Eurospeech 99*, pages 1151–1154.

H. Hastie, M. Johnston, and P. Ehlen. 2002. Context-sensitive Help for Multimodal Dialogue. In *Proceedings of the 4th IEEE International Conference on Multimodal Interfaces*, pages 93–98, Pittsburgh, PA.

M. Johnston and S. Bangalore. 2000. Finite-state Multimodal Parsing and Understanding. In *Proceedings of COLING 2000*, pages 369–375, Saarbrücken, Germany.

M. Johnston, S. Bangalore, and G. Vasireddy. 2001. MATCH: Multimodal Access To City Help. In *Workshop on Automatic Speech Recognition and Understanding*, Madonna di Campiglio, Italy.

M. Johnston, S. Bangalore, A. Stent, G. Vasireddy, and P. Ehlen. 2002a. Multimodal Language Processing for Mobile Information Access. In *Proceedings of ICSLP 2002*, pages 2237–2240.

M. Johnston, S. Bangalore, G. Vasireddy, A. Stent, P. Ehlen, M. Walker, S. Whittaker, and P. Maloor. 2002b. MATCH: An Architecture for Multimodal Dialog Systems. In *Proceedings of ACL-02*, pages 376–383.

L. Lamel, S. Bennacef, J. L. Gauvain, H. Dartigues, and J. N. Temem. 2002. User Evaluation of the MASK Kiosk. *Speech Communication*, 38(1-2):131–139.

S. Narayanan, G. DiFabbrizio, C. Kamm, J. Hubbell, B. Buntschuh, P. Ruscitti, and J. Wright. 2000. Effects of Dialog Initiative and Multi-modal Presentation Strategies on Large Directory Information Access. In *Proceedings of ICSLP 2000*, pages 636–639.

S. Oviatt and P. Cohen. 2000. Multimodal Interfaces That Process What Comes Naturally. *Communications of the ACM*, 43(3):45–53.

R. Raisamo. 1998. A Multimodal User Interface for Public Information Kiosks. In *Proceedings of PUI Workshop, San Francisco*.

W. Wahlster. 2003. SmartKom: Symmetric Multimodality in an Adaptive and Reusable Dialogue Shell. In R. Krahl and D. Gunther, editors, *Proceedings of the Human Computer Interaction Status Conference 2003*, pages 47–62.