TELEMORPH: BANDWIDTH-DETERMINED MOBILE MULTIMODAL PRESENTATION

ANTHONY SOLON, PAUL McKEVITT, and KEVIN CURRAN

Intelligent Multimedia Research Group, School of Computing and Intelligent Systems, Faculty of Engineering, University of Ulster, Magee Campus, Northland Road, Northern Ireland, BT48 7JL, UK

This article presents the initial stages of research at the University of Ulster into a mobile intelligent multimedia presentation system called TeleMorph. TeleMorph aims to dynamically generate multimedia presentations using output modalities that are determined by the bandwidth available on a mobile device's wireless connection. To demonstrate the effectiveness of this research, TeleTuras, a tourist information guide for the city of Derry, will implement the solution provided by TeleMorph, thus demonstrating its effectiveness. This article does not focus on the multimodal content composition but rather concentrates on the motivation for and issues surrounding such intelligent tourist systems.

Key words: Mobile intelligent multimedia; Intelligent multimedia generation and presentation; Intelligent tourist interfaces

Introduction

Whereas traditional interfaces support sequential and unambiguous input from keyboards and conventional pointing devices (e.g., mouse, trackpad), intelligent multimodal interfaces relax these constraints and typically incorporate a broader range of input devices [e.g., spoken language, eye and head tracking, three dimensional (3D) gesture] (Maybury, 1999). The integration of multiple modes of input as outlined by Maybury allows users to benefit from the optimal way in which human communication works. Although humans have a natural facility for managing and exploiting multiple input and output media, computers do not. To incorporate multimodality in user interfaces enables computer behavior to become analogous to human communication paradigms, and therefore the interfaces are easier to learn and use. Because there are large individual differences in ability and preference to use different modes of communication, a multimodal interface permits the user to exercise selection and control over how they interact with the computer (Fell et al., 1994). In this respect, multimodal inter-

Address correspondence to Kevin Curran, Intelligent Multimedia Research Group, School of Computing and Intelligent Systems, Faculty of Engineering, University of Ulster, Magee Campus, Northland Road, Northern Ireland, BT48 7JL, UK. Tel: +44 (028) 7137 5565; Fax: +44 (028) 7137 5470; E-mail: kj.curran@ulster.ac.uk

faces have the potential to accommodate a broader range of users than traditional graphical user interfaces (GUIs) and unimodal interfaces—including users of different ages, skill levels, native language status, cognitive styles, sensory impairments, and other temporary or permanent handicaps or illnesses.

Interfaces involving spoken or pen-based input, as well as the combination of both, are particularly effective for supporting mobile tasks, such as communications and personal navigation. Unlike the keyboard and mouse, both speech and pen are compact and portable. When combined, people can shift these input modes from moment to moment as environmental conditions change (Holzman, 1999). Implementing multimodal user interfaces on mobile devices is not as clear-cut as doing so on ordinary desktop devices. This is due to the fact that mobile devices are limited in many respects: memory, processing power, input modes, battery power, and an unreliable wireless connection with limited bandwidth. This project researches and implements a framework for multimodal interaction in mobile environments taking into consideration fluctuating bandwidth. The system output is bandwidth dependent, with the result that output from semantic representations is dynamically morphed between modalities or combinations of modalities. With the advent of 3G wireless networks and the subsequent increased speed in data transfer available, the possibilities for applications and services that will link people throughout the world who are connected to the network will be unprecedented. One may even anticipate a time when the applications and services available on wireless devices will replace the original versions implemented on ordinary desktop computers. Some projects have already investigated mobile intelligent multimedia systems, using tourism in particular as an application domain. Koch (2000) is one such project, which analyzed and designed a position-aware speech-enabled hand-held tourist information system for Aalborg in Denmark. This system is position and direction aware and uses these abilities to guide a tourist on a sightseeing tour. In TeleMorph bandwidth will primarily determine the modality/modalities utilized in the output presentation, but also factors such as device constraints, user goal, and user situationalization will be taken into consideration. A provision will also be integrated that will allow users to choose their preferred modalities.

The main point to note about these systems is that current mobile intelligent multimedia systems fail to take into consideration network constraints and especially the bandwidth available when transforming semantic representations into the multimodal output presentation. If the bandwidth available to a device is low, then it is obviously inefficient to attempt to use video or animations as the output on the mobile device. This would result in an interface with depreciated quality, effectiveness, and user acceptance. This is an important issue as regards the usability of the interface. Learnability, throughput, flexibility, and user attitude are the four main concerns affecting the usability of any interface. In the case of the previously mentioned scenario (reduced bandwidth = slower/inefficient output), the throughput of the interface is affected and as a result the user's attitude is also. This is only a problem when the required bandwidth for the output modalities exceeds that which is available, hence the importance of choosing the correct output modality/modalities in relation to available resources.

Background and Related Work

Elting, Zwickel, and Malaka (2002) explain the cognitive load theory where two separate subsystems for visual and auditory memory work relatively independently. The load can be reduced when both subsystems are active, compared to processing all information in a single subsystem. Due to this reduced load, more resources are available for processing the information in more depth and thus for storing in long-term memory. This theory, however, only holds when the information presented in different modalities is not redundant, otherwise the result is an increased cognitive load. If, however, multiple modalities are used, more memory traces should be available (e.g., memory traces for the information presented auditorily and visually) even though the information is redundant, thus counteracting the effect of the higher cognitive load. Elting et al. investigated the effects of display size, device type, and style of multimodal presentation on working memory load, effectiveness for human information processing, and user acceptance. The aim of this research was to discover how different physical output devices affect the user's way of working with a presentation system, and to derive presentation rules

from this that adapt the output to the devices the user is currently interacting with. They intended to apply the results attained from the study in the EMBASSI project where a large set of output devices and system goals have to be dealt with by the presentation planner. Accordingly, they used a desktop PC, TV set with remote control, and a PDA as presentation devices, and investigated the impact the multimodal output of each of the devices had on the users. As a gauge, they used the recall performance of the users on each device. The output modality combination for the three devices consisted of:

- plain graphical text output (T),
- text output with synthetic speech output of the same text (TS),
- a picture together with speech output (PS),
- graphical text output with a picture of the attraction (TP),
- graphical text, synthetic speech output, and a picture in combination (TPS).

The results of their testing on PDAs are relevant to any mobile multimodal presentation system that aims to adapt the presentation to the cognitive requirements of the device. Figure 1a shows the presentation appeal of various output modality combinations on various devices and Figure 1b shows mean recall performance of various output modality combination outputs on various devices.

The results show that in the TV and PDA group the PS combination proved to be the most efficient (in terms of recall) and second most efficient for desktop PC. So pictures plus speech appear to be a very convenient way to convey information to the user on all three devices. This result is theoretically supported by Baddeley's "Cognitive Load Theory" (Baddeley & Logie 1999, Sweller, van Merrienboer, & Paas, 1998), which states that PS is a very efficient way to convey information by virtue of the fact that the information is processed both auditorily and visually but with a moderate cognitive load. Another phenomenon that was observed was that the decrease of recall performance in time was especially significant in the PDA group. This can be explained by the fact that the work on a small PDA display resulted in a high cognitive load. Due to this load, recall performance decreased significantly over time. With respect to presentation appeal, it was not the most efficient modality combination that proved to be the most appealing (PS) but a combination involving a rather high cognitive load, namely TPS. The study showed that cognitive overload is a serious issue in user interface design, especially on small mobile devices. From their testing, Elting et al. discovered that when a system wants to present data to the user that is important to be remembered (e.g., a city tour), the most effective presentation mode should be used (picture & speech), which does not cognitively overload the user. When the system simply has to inform the user (e.g., about an interesting sight nearby) the most appealing/accepted presentation mode should be used (picture, text & speech). These points should be incorporated into multimodal presentation systems to achieve ultimate usability. This theory will be used in TeleMorph in the decision-making process, which determines what combinations of modalities are best suited to the current situation when designing the output presentation; that is, whether the system is presenting information that is important to be remembered (e.g., directions) or that is just informative (e.g., information on a tourist site).

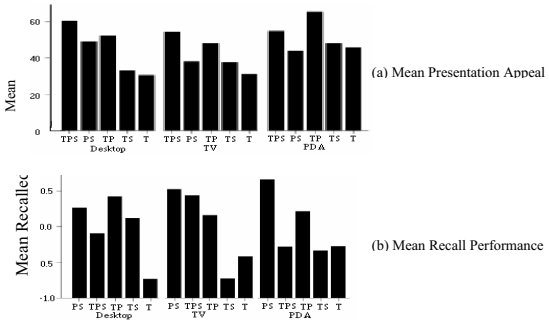


Figure 1. The most effective and most acceptable modality combinations.

TeleMorph

The aim of the TeleMorph project is to create a system that dynamically morphs between output modalities depending on available network bandwidth. The aims are to:

- Determine a wireless system's output presentation (unimodal/multimodal) depending on the

network bandwidth available to the mobile device connected to the system.
- Implement TeleTuras, a tourist information guide for the city of Derry (Northern Ireland), and integrate the solution provided by TeleMorph, thus demonstrating its effectiveness.

The aims entail the following objectives, which include receiving and interpreting questions from the user; mapping questions to multimodal semantic representation; matching multimodal representation to database to retrieve answer; mapping answers to multimodal semantic representation; querying bandwidth status and generating multimodal presentation based on bandwidth data. The domain chosen as a test bed for TeleMorph is *e*Tourism. The system to be developed, called TeleTuras, is an interactive tourist navigation aid for tourists in the city of Derry. It will incorporate route planning, maps, points of interest, spoken presentations, graphics of important objects in the area, and animations. The main focus will be on the output modalities used to communicate this information and also the effectiveness of this communication. The tools that will be used to implement this system are detailed in the next section. TeleTuras will be capable of taking input queries in a variety of modalities whether they are combined or used individually. Queries can also be directly related to the user's position and movement direction, enabling questions/commands such as:

- "Where is the Leisure Center?"
- "Take me to the Council Offices"
- "What buildings are of interest in this area?" (While circling a certain portion of the map on the mobile device, or perhaps if the user wants information on buildings of interest in their current location, they need not identify a specific part of the map as the system will wait until the timing threshold is passed and then presume no more input modalities relating to this inquiry).

Java 2 Micro Edition (J2ME) is an ideal programming language for developing TeleMorph, as it is the target platform for the Java Speech API (JSAPI) (Java Community Process, http://www.jcp.org/en/home/index). The JSAPI enables the inclusion of speech technology in user interfaces for Java applets and applications. The Java Speech API Markup Language (JSML, 2002) and the Java Speech API Grammar Format (JSGF, 2002) are companion specifications to the JSAPI. JSML (currently in beta) defines a standard text format for marking up text for input to a speech synthesizer. JSGF version 1.0 defines a standard text format for providing a grammar to a speech recognizer. JSAPI does not provide any speech functionality itself, but through a set of APIs and event interfaces, access to speech functionality provided by supporting speech vendors is accessible to the application. As it is inevitable that a majority of tourists will be foreigners, it is necessary that TeleTuras can process multilingual speech recognition and synthesis. To support this, an IBM implementation of JSAPI "speech for Java" will be utilized. It supports US and UK English, French, German, Italian, Spanish, and Japanese. To incorporate the navigation aspect of the proposed system a positioning system is required. The global positioning system (GPS) (Koch, 2000) will be employed to provide the accurate location information necessary for a location-based service (LBS). The user interface (UI) defined in J2ME is logically composed of two sets of APIs: high-level UI API, which emphasizes portability across different devices, and low-level UI API, which emphasizes flexibility and control. TeleMorph will use a dynamic combination of these in order to provide the best solution possible. An overview of the architecture to date is shown in Figure 2.

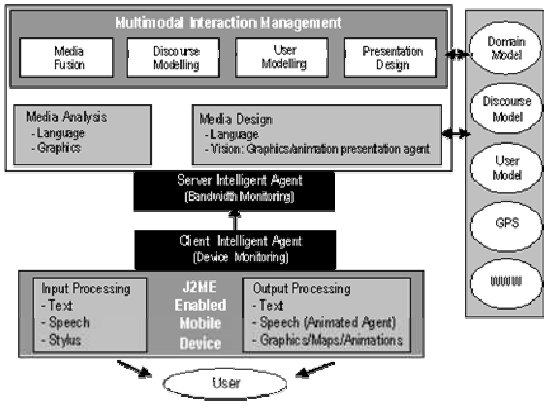


Figure 2. TeleMorph architecture.

Media Design takes the output information and morphs it into relevant modality/modalities depending on the information it receives from the Server Intelligent Agent regarding available bandwidth, while also taking into consideration the Cognitive Load Theory as described earlier. Media Analysis receives input from the Client Device and analyzes it to distinguish the modality types that the user utilized in their input. The Domain Model, Discourse Model, User Model, GPS, and WWW are additional sources of information for the Multimodal Interaction Manager, which assist it in producing an appropriate and correct output presentation. The Server Intelligent Agent is responsible for monitoring bandwidth, sending streaming media that is morphed to the appropriate modalities, and receiving input from Client Device & mapping to Multimodal Interaction Manager. The Client Intelligent Agent is in charge of monitoring device constraints (e.g., memory available, sending multimodal information on input to the server, and receiving streamed multimedia).

Data Flow of TeleMorph

The data flow within TeleMorph is shown in Figure 3, which details the data exchange among the main components. Figure 3 shows the flow of control in TeleMorph. The *Networking API* sends all input from the client device to the TeleMorph server. Each time this occurs, the *Device Monitoring* module will retrieve information on the client device's status and this information is also sent to the server. On input the user can make a multimodal query to the system to stream a new presentation which will consist of media pertaining to their specific query. TeleMorph will receive requests in the *Interaction Manager* and will process requests via the *Media Analysis* module, which will pass semantically useful data to the *Constraint Processor* where modalities suited to the current network bandwidth (and other constraints) will be chosen to represent the information. The presentation is then designed using these modalities by the *Presentation Design* module. The media are processed by the *Media Allocation* module and following this the complete multimodal Synchronized Multimedia Integration Language (SMIL) (Rutledge, 2001) presentation is passed to the *Streaming Server* to be streamed to the client device. A user can also input particular modality/cost choices on the TeleMorph client. In this way the user can morph the current presentation they are receiving to a presentation consisting of specific modalities that may be better suited their current situation (driving/walking) or environment (work/class/pub). This path through TeleMorph is identified by the dotted line in Figure 3. Instead of analyzing and interpreting the media, TeleMorph simply stores these choices using the *User Prefs* module and then redesigns the presentation as normal using the *Presentation Design* module. The *Media Analysis* module that passes semantically useful data to the *Constraint Processor* consists of lower level elements that are portrayed in Figure 4. As can be seen, the input from the user is processed by the *Media Analy-*

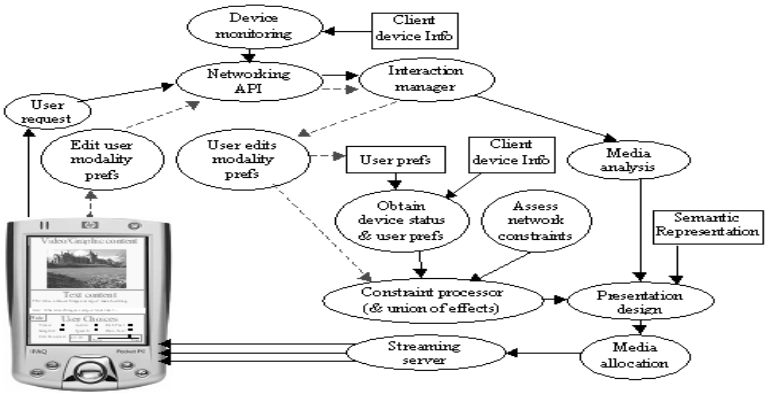


Figure 3. TeleMorph flow of control.

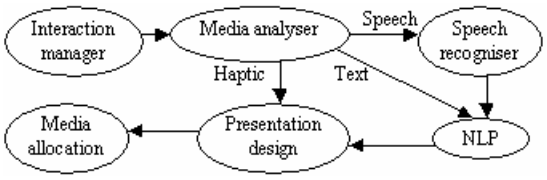


Figure 4. Media analysis data flow.

sis module, identifying Speech, Text and Haptic modalities.

The speech needs to be processed initially by the speech recognizer and then interpreted by the *NLP* module. Text also needs to be processed by the *NLP* module in order to attain its semantics. Then the *Presentation Design* module takes these input modalities and interprets their meaning as a whole and designs an output presentation using the semantic representation. This is then processed by the *Media Allocation* modules.

The Mobile Client's Output Processing module will process media being streamed to it across the wireless network and present the received modalities to the user in a synchronized fashion. The Input Processing module on the client will process input from the user in a variety of modes. This module will also be concerned with timing thresholds between different modality inputs. In order to implement this architecture for initial testing, a scenario will be set up where switches in the project code will simulate changing between a variety of bandwidths. To implement this, TeleMorph will draw on a database that will consist of a table of bandwidths ranging from those available in 1G, 2G, 2.5G (GPRS), and 3G networks. Each bandwidth value will have access to related information on the modality/combinations of modalities that can be streamed efficiently at that transmission rate. The modalities available for each of the aforementioned bandwidth values (1G–3G) will be worked out by calculating the bandwidth required to stream each modality (e.g., text, speech, graphics, video, animation). Then the amalgamations of modalities that are feasible are computed.

Client Output

Output on thin client devices connected to TeleMorph will primarily utilize a SMIL media player that will present video, graphics, text, and speech to the end user of the system. The J2ME Text-To-Speech (TTS) engine processes speech output to the user. An autonomous agent will be integrated into the TeleMorph client for output as they serve as an invaluable interface agent to the user as they incorporate modalities that are the natural modalities of face-to-face communication among humans. A SMIL media player will output audio on the client device. This audio will consist of audio files that are streamed to the client when the necessary bandwidth is available. However, when sufficient bandwidth is unavailable audio files will be replaced by ordinary text that will be processed by a TTS engine on the client producing synthetic speech output.

Autonomous Agents in TeleTuras

An autonomous agent will serve as an interface agent to the user as they incorporate modalities that are the natural modalities of face-to-face communication among humans. It will assist in communicating information on a navigation aid for tourists about sites, points of interest, and route planning. Microsoft Agent (http://www.microsoft.com./msagent/default.asp) provides a set of programmable software services that supports the presentation of interactive animated characters. It enables developers to incorporate conversational interfaces, which leverage natural aspects of human social communication. In addition to mouse and keyboard input, Microsoft Agent includes support for speech recognition so applications can respond to voice commands. Characters can respond using synthesized speech, recorded audio, or text. One advantage of agent characters is they provide higher levels of a character's movements often found in the performance arts, like blink, look up, look down, and walk. BEAT, another animator's tool that was incorporated in Real Estate Agent (REA) (Cassell, Sullivan, Prevost, & Chruchill, 2000) allows animators to input typed text that they wish to be spoken by an animated figure. These tools can all be used to implement actors in TeleTuras.

Client Input

The TeleMorph client will allow for speech recognition, text, and haptic deixis (touch screen) input. A speech recognition engine will be reused to

process speech input from the user. Text and haptic input will be processed by the J2ME graphics API. Speech recognition in TeleMorph resides in *Capture Input* as illustrated in Figure 5.

The Java Speech API Mark-up Language (http://java.sun.com/products/java-media/speech/) defines a standard text format for marking up text for input to a speech synthesizer. As mentioned before, JSAPI does not provide any speech functionality itself, but through a set of APIs and event interfaces, access to speech functionality (provided by supporting speech vendors) is accessible to the application. For this purpose IBM's implementation of JSAPI "speech for Java" is adopted for providing multilingual speech recognition functionality. This implementation of the JSAPI is based on ViaVoice, which will be positioned remotely in the *Interaction Manager* module on the server. The relationship between the JSAPI speech recognizer (in the *Capture Input* module in Fig. 5) on the client and ViaVoice (in the *Interaction Manager* module in Fig. 5) on the server is necessary as speech recognition is computationally too heavy to be processed on a thin client. After the ViaVoice speech recognizer has processed speech, which is input to the client device, it will also need to be analyzed by an *NLP* module to assess its semantic content. A reusable tool to do this is yet to be decided upon to complete this task. Possible solutions for this include adding an additional NLP component to ViaVoice, or perhaps reusing other natural understanding tools such as PC-PATR (McConnel, 1996), which is a natural language parser based on context-free phrase structure grammar and unifications on the feature structures associated with the constituents of the phrase structure rules.

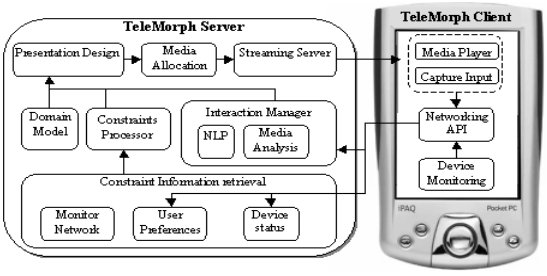


Figure 5. Modules within TeleMorph.

Graphics

The UI defined in J2ME is logically composed of two sets of APIs: high-level UI API, which emphasizes portability across different devices, and low-level UI API, which emphasizes flexibility and control. The portability in the high-level API is achieved by employing a high level of abstraction. The actual drawing and processing user interactions are performed by implementations. Applications that use the high-level API have little control over the visual appearance of components, and can only access high-level UI events. On the other hand, using the low-level API, an application has full control of appearance, and can directly access input devices and handle primitive events generated by user interaction. However the low-level API may be device dependent, so applications developed using it will not be portable to other devices with a varying screen size. TeleMorph uses a combination of these to provide the best solution possible. Using these graphics APIs, TeleMorph implements a *Capture Input* module that accepts text from the user. Also using these APIs, haptic input is processed by the *Capture Input* module to keep track of the user's input via a touch screen, if one is present on the device. User preferences in relation to modalities and cost incurred are managed by the *Capture Input* module in the form of standard check boxes and text boxes available in the J2ME high-level graphics API.

Networking

Networking takes place using sockets in the *J2ME Networking API* module as shown in Figure 5 to communicate data from the *Capture Input* module to the *Media Analysis* and *Constraint Information Retrieval* modules on the server. Information on client device constraints will also be received from the *Device Monitoring* module to the *Networking API* and sent to the relevant modules within the *Constraint Information Retrieval* module on the server. Networking in J2ME has to be very flexible to support a variety of wireless devices and has to be device specific at the same time. To meet this challenge, the Generic Connection Framework (GCF) is incorporated into J2ME. The idea of the GCF is to define the abstractions of the networking and file input/output as generally as possible to support a broad range of devices, and leave the actual imple-

mentations of these abstractions to the individual device manufacturers. These abstractions are defined as Java interfaces. The device manufacturers choose which one to implement based on the actual device capabilities.

Client Device Status

A SysInfo J2ME application (or MIDlet) is used for easy retrieval of a device's capabilities in the *Device Monitoring* module as shown in Figure 5. It probes several aspects of the J2ME environment it is running in and lists the results. In particular, it tries to establish a networking connection to find out which protocols are available, check device memory, the Record Management System (RMS), and other device properties. The following are an explanation of the various values collected by the MIDlet.

- **Properties:** Contains basic properties that can be queried via System.getProperty(). They reflect the configuration and the profiles implemented by the device as well as the current locale and the character encoding used. The *platform* property can be used to identify the device type, but not all vendors support it.
- **Memory:** Displays the total heap size that is available to the Java virtual machine as well as the flash memory space available for RMS. The latter value will depend on former RMS usage of other MIDlets in most cases, so it doesn't really reflect the total RMS space until you run SysInfo on a new or "freshly formatted" MIDP device. The MIDlet also tries to detect whether the device's garbage collector is compacting (i.e., whether it is able to shift around used blocks on the heap to create one large block of free space instead of a large number of smaller ones).
- **Screen:** Shows some statistics for the device's screen, most notably the number of colors or grayscales and the resolution. The resolution belongs to the canvas that is accessible to MIDlets, not to the total screen, because the latter value can't be detected.
- **Protocols:** Lists the protocols that are supported by the device. HTTP is mandatory according to the J2ME MIDP specification, so this one should be available on every device. The other protocols are identified by the prefix used for them in the *Connector* class such as http—Hypertext Transfer Protocol (HTTP), https—Secure Hypertext Transfer Protocol (HTTPS), socket—Plain Transmission Control Protocol (TCP), ssocket—Secure Transmission Control Protocol (TCP+TLS), and serversocket—allows to listen in incoming connections (TCP) among others.
- **Limits:** Reflects some limitations that a device has. Most devices restrict the maximum length of the TextField and TextBox classes to 128 or 256 characters. Trying to pass longer contents using the setString() method might result in an IllegalArgumentException being thrown, so it is best to know these limitations in advance and work around them. Also, several devices limit the total number of record stores, the number of record stores that can be open at the same time, and the number of concurrently open connections. For all items, "none" means that no limit could be detected.
- **Speed:** The MIDlet also does some benchmarking for RMS access and overall device speed. This last section holds values gained during these benchmarks. The first four items show the average time taken for accessing an RMS record of 128 bytes using the given method. The last item shows the time it took the device to calculate the first 1000 prime numbers using a straightforward implementation of Eratosthenes' prime sieve algorithm. While this is not meant to be an accurate benchmark of the device's processor, it can give an impression of the general execution speed (or slowness) of a device and might be a good hint when to include a "Please wait" dialog.

TeleMorph Server-Side

SMIL is utilized to form the semantic representation language in TeleMorph and will be processed by the *Presentation Design* module in Figure 5. The HUGIN development environment allows TeleMorph to develop its decision-making process using Causal Probabilistic Networks, which will form the *Constraint Processor* module as portrayed

in Figure 5. The ViaVoice speech recognition software resides within the *Interaction Manager* module. On the server end of the system Darwin streaming server (http://developer.apple.com/Darwin/projects/darwin/) is responsible for transmitting the output presentation from the TeleMorph server application to the client device's *Media Player*.

SMIL Semantic Representation. The XML based Synchronised Multimedia Integration Language (SMIL) language (Rutledge, 2001) forms the semantic representation language of TeleMorph used in the *Presentation Design* module as shown in Figure 5. TeleMorph designs SMIL content that comprises multiple modalities that exploit currently available resources fully, while considering various constraints that affect the presentation, but in particular, bandwidth. This output presentation is then streamed to the *Media Player* module on the mobile client for displaying to the end user. TeleMorph will constantly recycle the presentation SMIL code to adapt to continuous and unpredictable variations of physical system constraints (e.g., fluctuating bandwidth, device memory), user constraints (e.g., environment), and user choices (e.g., streaming text instead of synthesized speech). In order to present the content to the end user, a SMIL media player needs to be available on the client device. A possible contender to implement this is MPEG-7, as it describes multimedia content using XML.

TeleMorph Reasoning: CPNs/BBNs. Causal Probabilistic Networks aid in conducting reasoning and decision making within the *Constraints Processor* module (Fig. 5). In order to implement Bayesian Networks in TeleMorph, the HUGIN 2003(http://www.hugin.com/) (Jensen & Jianming, 1995) development environment is used. HUGIN provides the necessary tools to construct Bayesian networks. When a network has been constructed, one can use it for entering evidence in some of the nodes where the state is known and then retrieve the new probabilities calculated in other nodes corresponding to this evidence. A Causal Probabilistic Network (CPN)/Bayesian Belief network (BBN) is used to model a domain containing uncertainty in some manner. It consists of a set of nodes and a set of directed edges between these nodes. A Belief Network is a Directed Acyclic Graph (DAG) where each node represents a random variable. Each node contains the states of the random variable it represents and a conditional probability table (CPT) or, in more general terms, a conditional probability function (CPF). The CPT of a node contains probabilities of the node being in a specific state given the states of its parents. Edges reflect cause–effect relations within the domain. These effects are normally not completely deterministic (e.g., disease $\rightarrow$ symptom). The strength of an effect is modeled as a probability.

JATLite Middleware. As TeleMorph is composed of several modules with different tasks to accomplish, the integration of the selected tools to complete each task is important. To allow for this a middleware is required within the *TeleMorph Server* as portrayed in Figure 5. One such middleware is JATLite (Jeon, Petrie, & Cutkosky, 2000), which was developed by the Stanford University. JATLite provides a set of Java packages that makes it easy to build multiagent systems using Java. Different layers are incorporated to achieve this, including:

- **Abstract layer:** Provides a collection of abstract classes necessary for JATLite implementation. Although JATLite assumes all connections to be made with TCP/IP, the abstract layer can be extended to implement different protocols such as User Datagram Protocol (UDP).
- **Base layer:** Provides communication based on TCP/IP and the abstract layer. There is no restriction on the message language or protocol. The base layer can be extended, for example, to allow inputs from sockets and output to files. It can also be extended to give agents multiple message ports.
- **Knowledge Query & Manipulation Language (KQML) layer:** Provides for storage and parsing of KQML messages and a router layer provides name registration/message routing and queuing for agents.

As an alternative to the JATLite middleware The Open Agent Architecture (OAA) (Cheyer & Martin, 2001) could be used. OAA is a framework for integrating a community of heterogeneous software agents in a distributed environment. Psyclone (2003, http://www.mindmakers.org/architectures.html) is a flexible middleware that can be used as a blackboard

server for distributed, multimodule and multiagent systems, which may also be utilized.

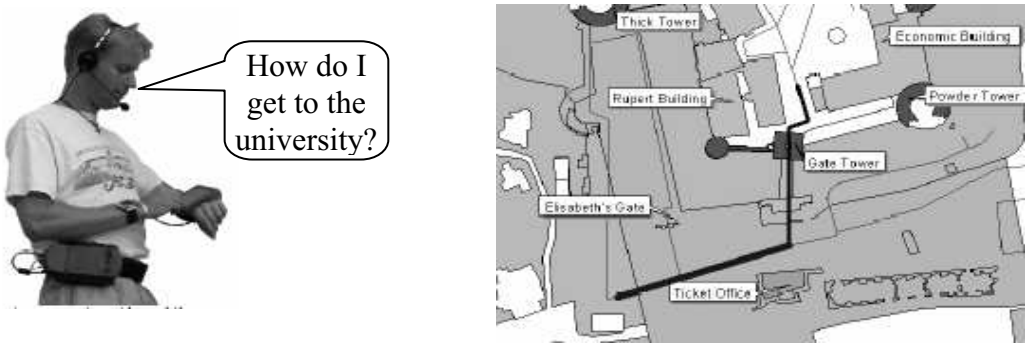
Related Work

SmartKom (Wahlster, 2001) is a multimodal dialogue system currently being developed by a consortium of several academic and industrial partners. The system combines speech, gesture, and facial expressions on the input and output side. The main scientific goal of SmartKom is to design new computational methods for the integration and mutual disambiguation of different modalities on a semantic and pragmatic level. SmartKom is a prototype system for a flexible multimodal human–machine interaction in two substantially different mobile environments: pedestrian and car. The system enables integrated trip planning using multimodal input and output. The key idea behind SmartKom is to develop a kernel system that can be used within several application scenarios. In a tourist navigation situation a user of SmartKom could ask a question about their friends who are using the same system (e.g., "Where are Tom and Lisa?" "What are they looking at?"). SmartKom is developing an XML-based mark-up language called MultiModal Markup Language (M3L) for the semantic representation of all of the information that flows between the various processing components. SmartKom is similar to TeleMorph and TeleTuras in that it strives to provide a multimodal information service to the end user. SmartKom-Mobile is specifically related to TeleTuras in the way it provides location-sensitive information of interest to the user of a thin client device about services or facilities in their vicinity.

DEEP MAP (Malaka, 2001; Malaka & Zipf, 2000) is a prototype of a digital personal mobile tourist guide that integrates research from various areas of computer science: geo-information systems, databases, natural language processing, intelligent user interfaces, knowledge representation, and more. The goal of DEEP MAP is to develop information technologies that can handle huge heterogeneous data collections, complex functionality, and a variety of technologies, but are still accessible for untrained users. DEEP MAP is an intelligent information system that may assist the user in different situations and locations providing answers to queries such as: Where am I? How do I get from A to B? What attractions are near by? Where can I find a hotel/restaurant? How do I get to the nearest Italian restaurant? The current prototype is based on a wearable computer called the Xybernaut. Examples of input and output in DEEP MAP are given in Figure 6a and b, respectively.

Figure 6a shows a user requesting directions to a university within their town using speech input. Figure 6b shows an example response to a navigation query. DEEP MAP displays a map that includes the user's current location and their destination, which are connected graphically by a line that follows the roads/streets interconnecting the two. Places of interest along the route are displayed on the map.

Other projects focusing on mobile intelligent multimedia systems, using tourism in particular as



(a) Example speech input (b) Example map

Figure 6. Example input and output in DEEP MAP.

an application domain, include Koch (2000) who describes one such project that analyzed and designed a position-aware speech-enabled hand-held tourist information system. The system is position and direction aware and uses these facilities to guide a tourist on a sight-seeing tour. (Rist, 2001) describes a system that applies intelligent multimedia to mobile devices. In this system a car driver can take advantage of online and offline information and entertainment services while driving. The driver can control phone and Internet access, radio, music repositories (DVD, CD-ROMs), navigation aids using GPS, and car reports/warning systems. (Pieraccini, 2002) outlines one of the main challenges of these mobile multimodal user interfaces, that being the necessity to adapt to different situations ("situationalization"). Situationalization as referred to by Pieraccini identifies that at different moments the user may be subject to different constraints on the visual and aural channels (e.g., walking while carrying things, driving a car, being in a noisy environment, wanting privacy, etc.).

EMBASSI (Hildebrand, 2000) explores new approaches for human–machine communication with specific reference to consumer electronic devices at home (TVs, VCRs, etc.), in cars (radio, CD player, navigation system, etc.), and in public areas (ATMs, ticket vending machines, etc.). Because it is much easier to convey complex information via natural language than by pushing buttons or selecting menus, the EMBASSI project focuses on the integration of multiple modalities like speech, haptic deixis (pointing gestures), and GUI input and output. Because EMBASSI's output is destined for a wide range of devices, the system considers the effects of portraying the same information on these different devices by utilizing CLT (Baddeley & Logie, 1999). Fink and Kobsa (2002) discuss a system for personalizing city tours with user modeling. They describe a user modeling server that offers services to personalized systems with regard to the analysis of user actions, the representation of the assumptions about the user, and the inference of additional assumptions based on domain knowledge and characteristics of similar users. Nemirovsky and Davenport (2002) describe a wearable system called GuideShoes, which uses aesthetic forms of expression for direct information delivery. GuideShoes utilizes music as an information medium and musical patterns as a means for navigation in an open space, such as a street. Cohen-Rose and Christiansen (2002) discuss a system called The Guide, which answers natural language queries about places to eat and drink with relevant stories generated by storytelling agents from a knowledge base containing previously written reviews of places and the food and drink they serve.

TeleMorph in Relation to Existing Intelligent Multimodal Systems

In the tables there are comparisons showing features of various mobile intelligent multimedia (Table 1) and intelligent multimedia systems (Table 2).

Table 1
Comparison of Mobile Intelligent Multimedia Systems

Systems	Device	Location Aware	Device Aware	User Aware	Cognitive Load Aware	Bandwidth Aware	Constraint Union
SmartKom	Compaq iPaq	X	X	X			
DEEP MAP	Xybernaut MA IV	X	X	X			
CRUMPET	unspecified mobile device	X		X			
VoiceLog	Fujitsu Stylistic 1200 pen PC		X	X			
MUST	Compaq iPaq	X					
Aalborg	Palm V	X					
GuideShoes	CutBrain CPU	X					
The Guide	mobile phone	X		X			
QuickSet	Fujitsu Stylistic 1000	X	X				
EMBASSI	consumer devices (e.g., navigation system)		X	X			
Pedersen & Larsen	Compaq iPaq	X	X	X			
TeleMorph	J2ME device		X	X	X	X	X

Table 2
Comparison of Intelligent Multimedia Systems

		NLP Component		Input Media				
Categories	Systems	Natural Language Generation	Natural Language Understanding	Text	Pointing (Haptic Deixis)	Speech	Vision	Text
Intelligent multimedia presentation systems	WIP		X	X				X
	COMET		X	X				X
	TEXTPLAN			X	X			X
	Cicero	X	X	X	X	X		X
	IMPROVISE			X				
Intelligent multimedia interfaces	AIMI			X	X	X		X
	AlFresco	X	X		X	X		X
	XTRA		X	X	X			X
	CUBRICON	X	X	X	X	X		
Mobile intelligent multimedia systems	SmartKom	X	X	X	X	X	X	X
	DEEP MAP		X	X	X	X		X
	CRUMPET	X			X	X		X
	VoiceLog		X		X	X		X
	MUST	X	X	X	X	X		X
	GuideShoes	X	X	X	X	X		
	The Guide	X	X	X	X	X		X
	QuickSet	X		X	X	X		X
Intelligent multimodal agents	Cassell's SAM & Rea (BEAT)	X	X	X		X	X	
	Gandalf		X				X	
This project	**TeleMorph & TeleTuras**	**X**	**X**	**X**	**X**	**X**		**X**

(Malaka, 2000) points out when discussing DEEP MAP that in dealing with handheld devices, "Resources such as power or networking bandwidth may be limited depending on time and location." From Table 1 it is clear that there are a wide variety of mobile devices being used in mobile intelligent multimedia systems. The issue of device diversity is considered by a number of the systems detailed in the table. Some of these systems are simply aware of the type of output device (e.g., PDA, desktop, laptop, TV—EMBASSI) and others are concerned with the core resources that are available on the client device (e.g., memory, CPU, output capabilities—SmartKom-mobile). Some of these systems also allow for some method of user choice/preference when choosing output modalities in order to present a more acceptable output presentation for the end user. Pedersen and Larsen (2003) describe a test system that analyzes the effect of user acceptance when output modalities are changed automatically or are changed manually by the user. This work is represented in Table 1 but no final system was developed as part of the project. One other factor that is relevant to mobile intelligent multimedia systems is CLT. This theory states the most efficient (judged by user retention) and the most appealing (user acceptable) modalities for portraying information on various types of devices (PDA, TV, desktop). One system that takes this theory into account is EMBASSI (Hildebrand, 2000). One main issue that the systems reviewed fail to consider is the effect imposed by the union of all the aforementioned constraints. Of the mobile intelligent multimedia systems in Table 1, some acknowledge that (1) network bandwidth and (2) device constraints are important issues, but most do not proceed to take these into consideration when mapping their semantic representation to an output presentation, as can be seen from the table. As can also be seen from Table 1, none of the currently available mobile intelligent multimedia systems design their output presentation relative to the amount of available bandwidth available on the wireless network connecting the device.

TeleMorph differs from these systems in that it is aware of all the constraints that have been mentioned. Primarily, TeleMorph is bandwidth aware in that it constantly monitors the network for fluctuations in the amount of data that can be transmitted per sec-

ond [measured in bits per second (bps)]. As mobile enabled devices vary greatly in their range of capabilities (CPU, memory available, battery power, input modes, screen resolution and color, etc.), TeleMorph is also aware of the constraints that exist on TeleMorph's client device and takes these into consideration when mapping to output presentation. TeleMorph is also be aware of user-imposed limitations, which will consist of the user's preferred modalities and a restriction set by them on the cost they will incur in downloading the presentation. One other factor that has been considered in designing the output for TeleMorph is CLT. TeleMorph uses CLT to assist in setting the output modalities for different types of information that are portrayed in a presentation, such as information that requires high levels of retention (e.g., a city tour) or information that calls for user acceptance (purely informative) oriented modalities (e.g., information about an interesting sight nearby). From Table 1 one can also identify that the combination of all these constraints as a union is also a unique approach. TeleMorph is aware of all the relevant constraints that a mobile multimodal presentation system should be concerned with. TeleMorph analyzes a union of these constraints and decides on the optimal multimodal output presentation. The method employed by TeleMorph to process these various constraints and utilize this information effectively to design the most suitable combinations of output modalities is the main challenge within this research project.

Table 2 shows that TeleMorph and TeleTuras utilize similar input and output modalities to those employed by other mobile intelligent multimedia presentation systems. (Please note that due to space restrictions, output modalities have been omitted.) One point to note about the intelligent multimedia presentation systems in Table 2 is that on input none of them integrate vision, while only one system uses speech and two use haptic deixis. In comparison, all of the mobile intelligent multimedia systems integrate speech and haptic deixis on input. Both Guide and Quickset use only text and static graphics as their output modalities, choosing to exclude speech and animation modalities. VoiceLog is an example of one of the mobile systems presented in Table 2 that does not include text input, allowing only for speech input. Hence, some of the systems in Table 2 fail to include some input and output modalities. VoiceLog (2002, http://www.bbn.com), MUST (Almeida et al., 2002), GuideShoes (Nemirovsky & Davenport, 2002), The Guide (Cohen-Rose & Christiansen, 2002), and QuickSet (Oviatt et al., 2000) all fail to include animation in their output. Of these, the latter three systems also fail to use speech on output. GuideShoes is the only other mobile intelligent multimedia system that outputs nonspeech audio, but this is not combined with other output modalities, so it could be considered a unimodal communication.

With TeleMorph's ability on the client side to receive a variety of streaming media/modalities, TeleTuras is able to present a multimodal output presentation including nonspeech audio that will provide relevant background music about a certain tourist point of interest (e.g., theater/concert venue). The focus with TeleMorph's output presentation lies in the chosen modalities and the rules and constraints that determine these choices. TeleMorph implements a comprehensive set of input and output modalities. On input TeleMorph handles text, speech, and haptic modalities, while output consists of text, TTS, nonspeech audio, static graphics, and animation. This provides output similar to that produced in most current intelligent multimedia systems that mix text, static graphics (including map, charts, and figures), and speech (some with additional nonspeech audio) modalities.

Conclusion

We have touched upon some aspects of mobile intelligent multimedia systems. Through an analysis of these systems a unique focus has been identified: "bandwidth-determined mobile multimodal presentation." This article has presented our proposed solution in the form of a mobile intelligent system called TeleMorph that dynamically morphs between output modalities depending on available network bandwidth. TeleMorph will be able to dynamically generate a multimedia presentation from semantic representations using output modalities that are determined by constraints that exist on a mobile device's wireless connection, the mobile device itself, and also those limitations experienced by the end user of the device. The output presentation will include language and vision modalities consisting of video, speech, nonspeech audio, and text. Input

to the system will be in the form of speech, text, and haptic deixis.

The objectives of TeleMorph are: (1) receive and interpret questions from the user, (2) map questions to multimodal semantic representation, (3) match multimodal representation to knowledge base to retrieve answer, (4) map answers to multimodal semantic representation, (5) monitor user preference or client side choice variations, (6) query bandwidth status, (7) detect client device constraints and limitations, and (8) generate multimodal presentation based on constraint data. The architecture, data flow, and issues in the core modules of TeleMorph such as constraint determination and automatic modality selection are also given.

Biographical Notes

Anthony Solon is a Ph.D. student at the University of Ulster. His interests to date include intelligent multimedia, mobile computing, and multimodal bandwidth-determined streaming.

Paul McKevitt is Chair in Intelligent MultiMedia at the School of Computing & Intelligent Systems, Faculty of Engineering, University of Ulster (Magee College), Derry (Londonderry), Northern Ireland. His primary research interests are in natural language processing (NLP), including the processing of pragmatics, beliefs, and intentions in dialogue. He is also interested in philosophy, MultiMedia, and the general area of artificial intelligence.

Kevin Curran (B.Sc., Ph.D.) is a Lecturer at the University of Ulster in Northern Ireland. He has written over 90 academic research papers on areas such as distributed computing, emerging trends within wireless ad-hoc networks, dynamic protocol stacks, and mobile systems.

References

Almeida, L., Amdal, I., Beires, N., Boualem, M., Boves, L., den Os, E., Filoche, P., Gomes, R., Knudsen, J. E., Kvale, K., Rugelbak, J., Tallec, C., & Warakagoda, N. (2002). The MUST guide to Paris—implementation and expert evaluation of a multimodal tourist guide to Paris. In *Proceedings of ISCA Tutorial and Research Workshop (ITRW) on Multi-Modal Dialogue in Mobile Environments* (IDS 2002) (pp. 49–51). Kloster Irsee, Germany, June 17–19.

Baddeley, A. D., & Logie, R. H. (1999). Working memory: The multiple-component model. In A. Miyake & P. Shah (Eds.), *Models of working memory: Mechanisms of active maintenance and executive control* (pp. 28–61). Cambridge: Cambridge University Press.

Cassell, J., Sullivan, J., Prevost, S., & Churchill, E. (2000). *Embodied conversational agents.* Cambridge, MA: MIT Press.

Cheyer, A., & Martin, D. (2001). The open agent architecture. *Journal of Autonomous Agents and Multi-Agent Systems, 4*(1), 143–148.

Cohen-Rose, A. L., & Christiansen, S. B. (2002). The hitchhiker's guide to the galaxy. In P. McKevitt, S. Ó Nualláin, & C. Mulvihill (Eds.), *Language vision and music* (pp. 55–66). Amsterdam: John Benjamins.

Elting, C., Zwickel, J., & Malaka, R. (2002). Device-dependant modality selection for user-interfaces. *International Conference on Intelligent User Interfaces. Intelligent User Interfaces*, San Francisco, CA, January 13–16.

Fell, H., Delta, H., Peterson, R., Ferrier, L., et al. (1994). Using the baby-babble-blanket for infants with motor problems. *Conference on Assistive Technologies (ASSETS'94)* (pp. 77–84), Marina del Rey, CA.

Fink, J., & Kobsa, A. (2002). User modeling for personalised city tours. *Artificial Intelligence Review, 18*(1), 33–74.

Hildebrand, A. (2000). EMBASSI: Electronic multimedia and service assistance. In *Proceedings IMC'2000* (pp. 50–59), November, Rostock-Warnemünde, Germany.

Holzman, T. G. (1999). Computer-human interface solutions for emergency medical care. *Interactions, 6*(3), 13–24.

Jensen, F. V., & Jianming, L. (1995). Hugin: A system for hypothesis driven data request. In A. Gammerman (Ed.), *Probabilistic reasoning and Bayesian belief networks* (pp. 109–124). London, UK: Alfred Waller ltd.

Jeon, H., Petrie, C., & Cutkosky, M. R. (2000). JATLite: A Java agent infrastructure with message routing. *IEEE Internet Computing, 4*(2), 87–96.

Koch, U. O. (2000). *Position-aware speech-enabled hand held tourist information system.* Semester 9 project report, Institute of Electronic Systems, Aalborg University, Denmark.

Malaka, R. (2001). Multi-modal interaction in private environments. *International Seminar on Coordination and Fusion in MultiModal Interaction*, Schloss Dagstuhl International Conference and Research Center for Computer Science, Wadern, Saarland, Germany, October 29–November 2.

Malaka, R., & Zipf, A. (2000). DEEP MAP—challenging IT research in the framework of a tourist information system. *Proceedings of ENTER 2000, 7th International Congress on Tourism and Communications Technologies in Tourism*, Barcelona. Wien/New York: Springer Computer Science.

Maybury, M. T. (1999). Intelligent user interfaces: An introduction. *Intelligent User Interfaces* (pp. 5–8), January 3–4, Los Angeles, CA.

McConnel, S. (1996). KTEXT and PC-PATR: Unification based tools for computer aided adaptation. In H. A. Black, A. Buseman, D. Payne, & G. F. Simons (Eds.), *Proceedings of the 1996 General CARLA Conference* (pp. 39–95), November 14–15. Waxhaw, NC/Dallas: JAARS and Summer Institute of Linguistics.

Nemirovsky, P., & Davenport, G. (2002). Aesthetic forms

of expression as information delivery units. In P. McKevitt, S. Ó Nualláin, & C. Mulvihill (Eds.), *Language vision and music* (pp. 255–270). Amsterdam: John Benjamins.

Oviatt, S. L., Cohen, P. R., Wu, L., Vergo, J., Duncan, E., Suhm, B., Bers, J., Holzman, T., Winograd, T., Landay, J., Larson J., & Ferro, D. (2000). Designing the user interface for multimodal speech and gesture applications: State-of-the-art systems and research directions. *Human Computer Interaction, 15*, 263–322.

Pedersen, J. S., & Larsen, S. R. (2003). *A pilot study in modality shifting due to changing network conditions.* M.Sc. thesis, Center for Person Communication, Aalborg University, Denmark.

Pieraccini, R. (2002, Summer). Wireless multimodal—the next challenge for speech recognition. *ELSNews*, ii.2.

Rist, T. (2001). Media and content management in an intelligent driver support system. *International Seminar on Coordination and Fusion in MultiModal Interaction*, Schloss Dagstuhl International Conference and Research Center for Computer Science, Wadern, Saarland, Germany, October 29–November 2. http://www.dfki.de/~wahlster/Dagstuhl_Multi_Modality/rist-dagstuhl.pdf

Rutledge, L. (2001, September–October). SMIL 2.0: XML for Web multimedia. *IEEE Internet Computing*, 78–84.

Sweller, J., van Merrienboer, J. J. G., & Paas, F. G. W. C. (1998). Cognitive architecture and instructional design. *Educational Psychology Review, 10*, 251–296.

Wahlster, W. N. (2001). SmartKom a transportable and extensible multimodal dialogue system. *International Seminar on Coordination and Fusion in MultiModal Interaction*, Schloss Dagstuhl Int Conference and Research Center for Computer Science, Wadern, Saarland, Germany, October 29–November 2.