Mobile Multimodal Dynamic Output Morphing Tourist Systems

Anthony Solon, Paul Mc Kevitt, Kevin Curran
Intelligent Multimedia Research Group
University of Ulster, Magee Campus, Northland Road, Northern Ireland, BT48 7JL, UK
Email: {aj.solon, p.mckevitt, kj.curran@ulster.ac.uk}

Abstract

TeleMorph dynamically generates 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 paper concentrates on the motivation for & issues surrounding intelligent tourist systems.

1 Introduction

The integration of multiple modes of input allows users to benefit from the optimal way in which human communication works. Whereas traditional interfaces support sequential and un-ambiguous 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) [1]. 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 behaviour to become analogous to human communication paradigms, and therefore the interfaces are easier to learn and use. Since 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 [2]. In this respect, multimodal interfaces 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 [3]. 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. [4] is one such project which analysed 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 sight seeing tour. In TeleMorph bandwidth will primarily determine the modality/modalities utilised in the output presentation, but also factors such as device constraints, user goal and user situationalisation will be taken into consideration. A provision will also be integrated which 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's 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 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.

2 Related Work

SmartKom [5] 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, namely 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 which 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 M3L (MultiModal Markup Language) 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 [6, 7] is a prototype of a digital personal mobile tourist guide which integrates research from various areas of computer science: geo-information systems, data bases, 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. DEEP MAP displays a map which includes the user's current location and their destination, which are connected graphically by a line which 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 an application domain include [4] who describes one such project which analysed 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. [8] outlines one of the main challenges of these mobile multimodal user interfaces, that being the necessity to adapt to different situations ("situationalisation"). Situationalisation 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 whilst carrying things, driving a car, being in a noisy environment, wanting privacy etc.).

EMBASSI [9] 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.). Since 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 utilising Cognitive Load Theory (CLT) [10]. [11] discuss a system for personalising city tours with user modelling. They describe a user modelling server that offers services to personalised 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. [12] describe a wearable system called GuideShoes which uses aesthetic forms of expression for direct information delivery. GuideShoes utilises music as an information medium and musical patterns as a means for navigation in an open space, such as a street.

3 Cognitive Load Theory

[13] explain the cognitive load theory where two separate sub-systems for visual and auditory memory work relatively independently. The load can be reduced when both sub-systems are active, compared to processing all information in a single sub-system. 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 auditorially 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 combinations 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. 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" [9, 14], which states that PS is a very efficient way to convey information by virtue of the fact that the information is processed both auditorally 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, i.e. whether the system is presenting information which is important to be remembered (e.g. directions) or which is just informative (e.g. information on a tourist site).

4 Telemorph

The focus 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 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 information aid. 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?".

J2ME (Java 2 Micro Edition) is an ideal programming language for developing TeleMorph, as it is the target platform for the Java Speech API (JSAPI) [15]. The JSAPI enables the inclusion of speech technology in user interfaces for Java applets and applications. The Java Speech API Markup Language [16] and the Java Speech API Grammar Format [16] are companion specifications to the JSAPI. JSML (currently in beta) defines a standard text format for marking up text for input to a speech synthesiser. JSGF version 1.0 defines a standard text format for providing a grammar to a speech recogniser. 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, whilst also taking into consideration the Cognitive Load Theory as described earlier. Media Analysis receives input from the Client device and analyses it to distinguish the modality types that the user utilised in their input. The Domain Model, Discourse Model, User Model, GPS and WWW are additional sources of information for the Multimodal Interaction Manager that assist it in producing an appropriate and correct output presentation. The Server Intelligent Agent is responsible for monitoring bandwidth, sending streaming media which 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.

4.1 Data Flow of TeleMorph

The data flow within TeleMorph is shown in figure 1 which 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 Synchronised Multimedia Integration Language (SMIL) [17] presentation is passed to the *Streaming Server* to be streamed to the client device.

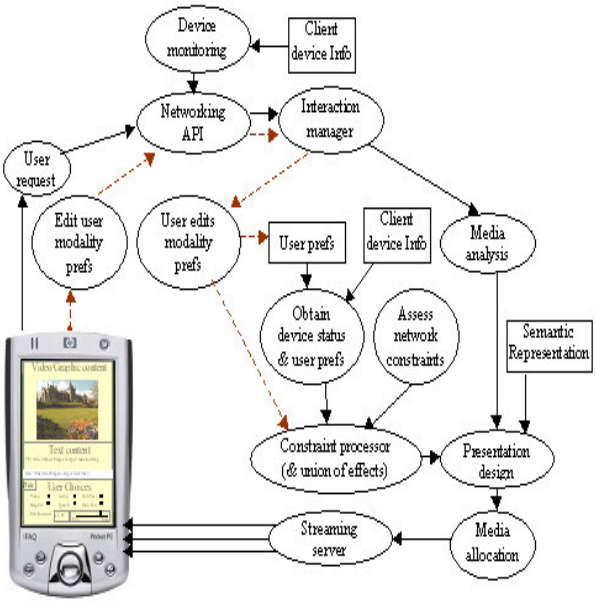

Figure 1: TeleMorph flow of control

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 which 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 1. Instead of analysing 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 2. As can be seen, the input from the user is processed by the *Media Analysis* module, identifying Speech, Text and Haptic modalities. The speech needs to be processed initially by the speech recogniser 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 synchronised 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.

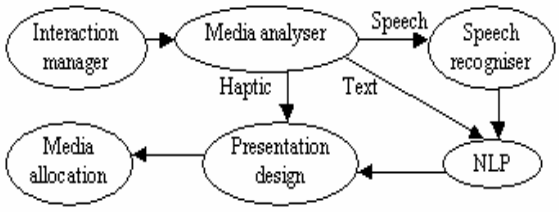

Figure 2: Media Analysis data flow

To implement this, TeleMorph will draw on a database which 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 fore-mentioned 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.

4.2 Client output

Output on thin client devices connected to TeleMorph will primarily utilise a SMIL media player which 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 which will be processed by a TTS engine on the client producing synthetic speech output.

4.3 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[1] 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 synthesised 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 which was incorporated in REA (Real Estate Agent) [18] 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.

4.4 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 3.

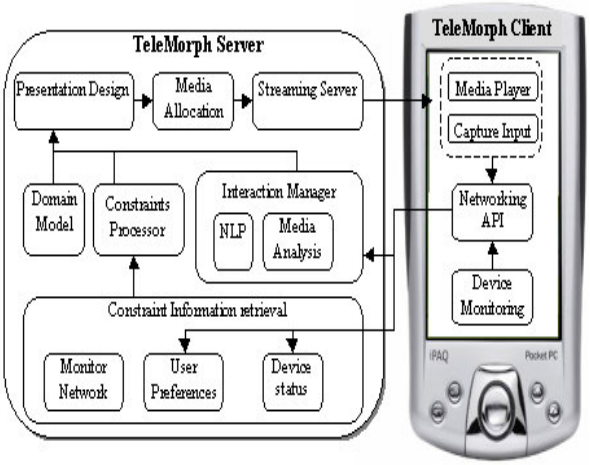

Figure 3: Modules within TeleMorph

The Java Speech API Mark-up Language[2] defines a standard text format for marking up text for input to a speech synthesiser. 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

[1] http://www.microsoft.com/msagent/default.asp

[2] http://java.sun.com/products/java-media/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 recogniser (in the *Capture Input* module in figure 5) on the client and ViaVoice (in the *Interaction Manager* module in figure 5) on the server is necessary as speech recognition is computationally too heavy to be processed on a thin client. After the ViaVoice speech recogniser has processed speech which is input to the client device, it will also need to be analysed 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 [19] 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.

4.5 Graphics

The User Interface (UI) defined in J2ME is logically composed of two sets of APIs, High-level UI API which emphasises portability across different devices and the Low-level UI API which emphasises 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 which 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.

4.6 TeleMorph Server-Side

SMIL is utilised 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[3] is responsible for transmitting the output presentation from the TeleMorph server application to the client device's *Media Player.*

4.6.1 SMIL semantic representation

The XML based Synchronised Multimedia Integration Language (SMIL) language [17] forms the semantic representation language of TeleMorph used in the *Presentation Design* module. TeleMorph designs SMIL content that comprises multiple modalities that exploit currently available resources fully, whilst 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 synthesised 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

[3] http://developer.apple.com/darwin/projects/darwin/

contender to implement this is MPEG-7, as it describes multimedia content using XML.

4.6.2 TeleMorph reasoning - CPNs/BBNs

Causal Probabilistic Networks aid in conducting reasoning and decision making within the *Constraints Processor* module (see figure 5). In order to implement Bayesian Networks in TeleMorph, the HUGIN [20] 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 -> symptom). The strength of an effect is modelled as a probability.

5 Future Work

Rather than trying to build TeleMorph from scratch, existing software tools will be made use of for speech recognition, Text-To-Speech (TTS), autonomous agent output, playing media presentations, attaining client device information, networking, bandwidth monitoring, Causal Probabilistic Networks (CPNs)/ Bayesian Belief Networks (BBNs) and streaming media. An analysis of reusable development tools for TeleMorph began at an early stage of the project and most have been decided upon.

5.1 SMIL media players

A *Media Player* will be used on the client side of TeleMorph to display the multimodal (animations, graphics, text, audio) presentation being received from the *Streaming Server*. It was decided to reuse a SMIL media player. Some SMIL 2.0 specification based players that are currently available include:

- AMBULANT Open Source SMIL Player by CWI.
- RealOne Platform by RealNetworks, which has full support for the SMIL 2.0 Language profile.
- GRiNS for SMIL-2.0 by Oratrix is a SMIL 2.0 player that supports SMIL 2.0 syntax and semantics.
- RubiC is developed by Roxia Co.,Ltd. It includes an authoring tool and player, and fully supports SMIL 2.0 specification. "RubiC" is also available for mobile handset for mobile internet MMS(Multimedia Messaging Service)
- TAO's announced Qi browser supports SMIL, HTML 4.01 CSS, and XML (including XML Parser, DTD and Schema validation).

Most of these SMIL players also include an edit/development tool for creating SMIL presentations. Various players have been identified and investigated in this project to determine their reusability and usefulness within the context of the TeleMorph architecture. Players will be analysed further to this to ensure compatibility during development of TeleMorph.

6 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 paper 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, non-speech audio and text. Input to the system will be in the form of speech, text and haptic deixis.

References

1. Maybury, M.T. (1999) Intelligent User Interfaces: An Introduction. *Intelligent User Interfaces*, 3-4 January 5-8, Los Angeles, California, USA.

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

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

[4] Koch, U.O. (2000) Position-aware Speech-enabled Hand Held Tourist Information System. *Semester 9 project report,* Institute of Electronic Systems, Aalborg University, Denmark.

[5] 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, 29 Oct-2 Nov

[6] Malaka, R. & A. Zipf (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 (Spain), Springer Computer Science, Wien, NY.

[7] 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, 29 October - 2 November.

8. Pieraccini, R., (2002) Wireless Multimodal – the Next Challenge for Speech Recognition. *ELSNews,* summer 2002, ii.2, Published by ELSNET, Utrecht, The Netherlands.

9. Hildebrand, A. (2000) EMBASSI: Electronic Multimedia and Service Assistance. *In Proceedings IMC'2000,* Rostock-Warnem¨unde, Germany, November, 50-59.

10. Baddeley, A. D. & R.H. Logie (1999) Working Memory: The Multiple-Component Model. In Miyake, A. and Shah, P. (Eds.), 28-61, Models of working memory: *Mechanisms of active maintenance and executive control*, Cambridge University Press.

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

12. Nemirovsky, P. & G. Davenport (2002) Aesthetic forms of expression as information delivery units. *In Language Vision and Music,* P. Mc Kevitt, S. Ó Nualláin and C. Mulvihill (Eds.), 255-270, Amsterdam: John Benjamins.

13. Elting, C., J. Zwickel & R. Malaka (2002) Device-Dependant Modality Selection for User-Interfaces, *International Conference on Intelligent User Interfaces. Intelligent User Interfaces*, San Francisco, CA, Jan 13-16, 2002.

14. Sweller, J., J.J.G. van Merrienboer & F.G.W.C. Paas (1998) Cognitive Architecture and Instructional Design. *Educational Psychology Review,* 10, 251-296.

15. JCP (2002) Java Community Process. http://www.jcp.org/en/home/index

16. JSML & JSGF (2002). Java Community Process. http://www.jcp.org/en/home/index Site visited 30/09/2003

17. Rutledge, L. (2001) SMIL 2.0: XML For Web Multimedia. *In IEEE Internet Computing*, Oct, 78-84.

18. Cassell, J., J. Sullivan, & E. Churchill, (2000) *Embodied Conversational Agents.* Cambridge, MA: MIT Press

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

20. HUGIN (2003) http://www.hugin.com/