# When talking is not enough[1]

Marianne Hickey
Mobile and Media Systems Laboratory
HP Laboratories Bristol
HPL-2002-171
May 2nd , 2003*

E-mail: marianne_hickey@hp.com

multimodal, voice, disability, pervasive interaction

The Internet of the future will be characterized by the availability of millions of e-services accessed by mobile users, via an array of different end-user devices, with a variety of user preferences and usage environments. How will people interact with these services and what technologies are required in the infrastructure and the devices? This article explores the future of user interaction with web services, in a world where people are increasingly mobile. I introduce one possible future - that of multimodal browsing, which offers you the choice of using spoken commands, keypads, pointing devices, gesture or other input modalities. On the output side, you have the choice of listening to spoken prompts and audio, or seeing plain text, motion video, or graphics. Interaction modalities can be used independently or concurrently, as is appropriate for the devices that are available, the environment and preferences. In this article, I discuss multimodal browsing and the research challenges and relate this to access to the web by people with disabilities.

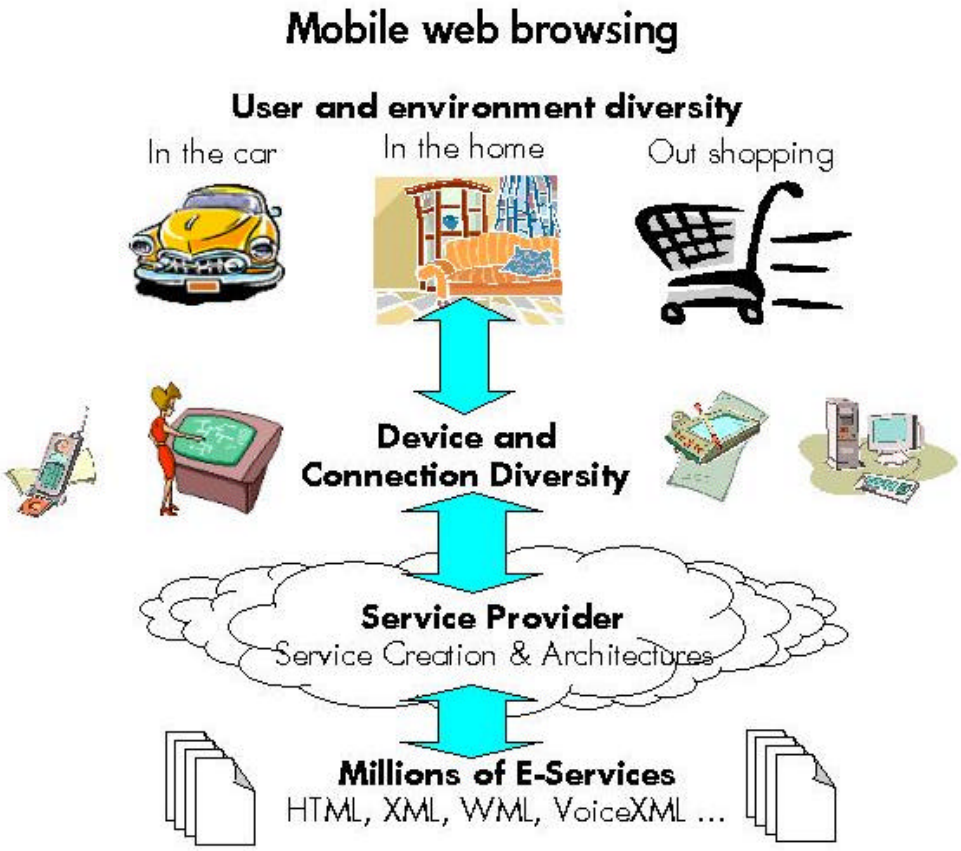
* Internal Accession Date Only
Approved for External Publication
[1] Ability Magazine, Issue 44 July 2002

# When talking is not enough

*Marianne Hickey*

HP Laboratories
Filton Road, Stoke Gifford, Bristol, BS34 8QZ, UK.
marianne_hickey@hp.com, http://www-uk.hpl.hp.com/

The Internet of the future will be characterized by the availability of millions of e-services accessed by mobile users, via an array of different end-user devices, with a variety of user preferences and usage environments. How will people interact with these services and what technologies are required in the infrastructure and the devices? This article explores the future of user interaction with web services, in a world where people are increasingly mobile. I introduce one possible future – that of multimodal browsing, which offers you the choice of using spoken commands, keypads, pointing devices, gesture or other input modalities. On the output side, you have the choice of listening to spoken prompts and audio, or seeing plain text, motion video, or graphics. Interaction modalities can be used independently or concurrently, as is appropriate for the devices that are available, the environment and preferences. In this article, I discuss multimodal browsing and the research challenges and relate this to access to the web by people with disabilities.

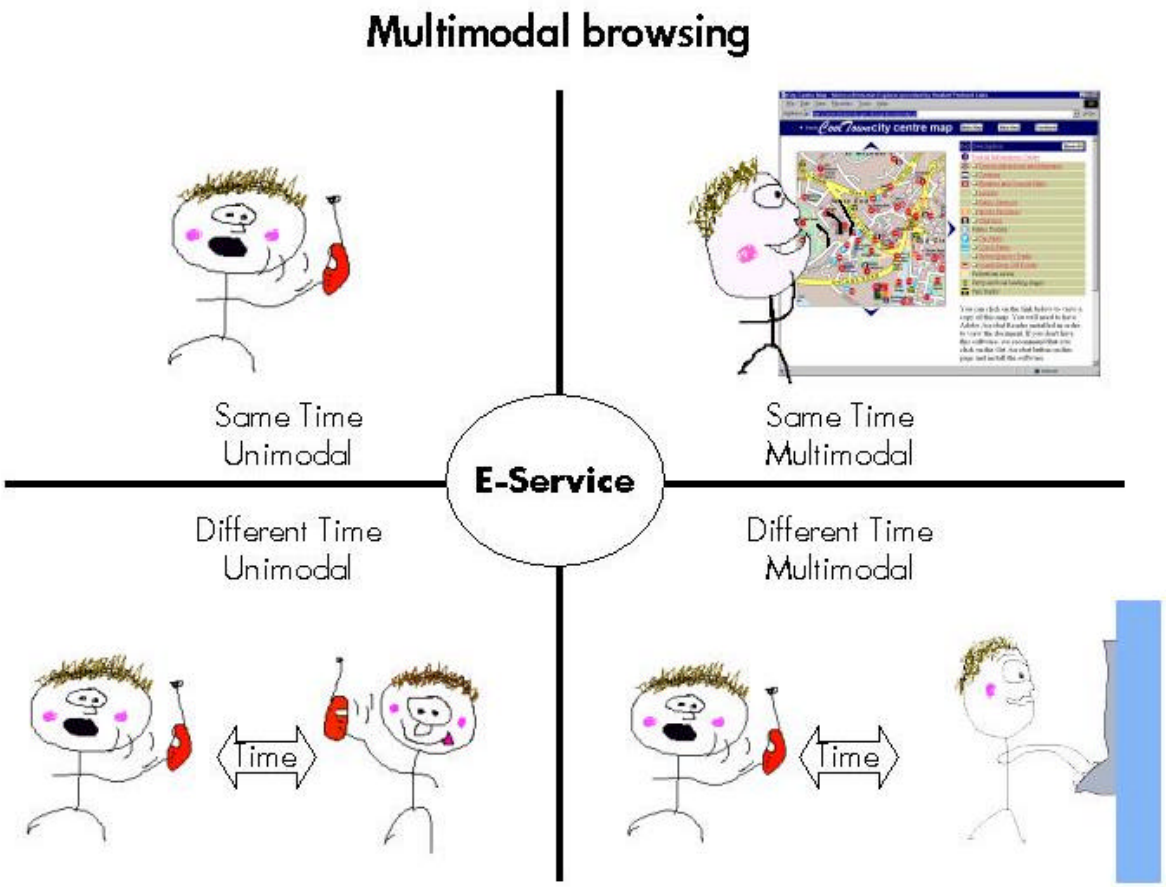


## *From voice to multimodal*

Voice is a key interaction modality for mobile users. Voice browsers present information using a combination of synthetic speech and pre-recorded audio, and allow the user to interact via spoken commands or phrases. This is great when hands and eyes are busy, as when you're driving. Voice interaction requires a small access device - a phone, for example - which many people already carry with them. The voice browser, speech recognition and synthesis can run in the network. Voice is well suited to when you need quick access to information and don't want to struggle with a tiny keypad or a stylus for character entry and menu navigation. While voice interaction has great promise, and although voice services are on the increase, more than voice is needed for truly mobile access to the web.

People use a variety of modalities to try to get a point across - for example, they can gesture, speak, draw, or handwrite - but current web browsers are either screen-based or voice-based. While voice is useful in many mobile situations, more flexible user interfaces are required that take advantage of multiple modalities. There are many situations where voice is not the most appropriate or the preferred interaction modality and using an alternative would help. For example, you really want to be able to switch to a visual interface when you need privacy, when it's noisy, or when it's more appropriate for the task - pointing to an item on a displayed list can be more comfortable than hearing the list and speaking a choice. And when you need a richer user interface, you can better communicate with multiple modalities: point to a location on a map and say, "Give me directions to here"; see an animation while hearing a weather forecast sent to your cell phone; see the lips, expressions, and gestures of your synthetic personal agent as you listen to your appointments for the day.

The vision of multimodal interaction is that as you move around you interact with a web service according to your context and preferences. You can move seamlessly between using different modalities or even different devices – for example, seeing or hearing, speaking, pointing, or typing. You may also set aside and re-visit services where you left off, perhaps using different access devices and interaction modalities. So multimodal interaction encompasses the choice of alternative modalities, the rich use of multiple modalities together, and also the ability to change to a different set of interaction modalities according to circumstance, perhaps via different end-user devices. That’s the vision we are working towards in our research at HP labs.




## *Multimodal browsing and disability*

I have so far talked about multimodal interfaces for the ever-increasing demand of mobility. There is another great potential of this kind of work, in terms of making the web more accessible to a wider range of people, including people with disabilities. Interaction with the web in a mobile setting poses many similar technical challenges to interaction with the web by people with disabilities. But this is not a one-way flow – technology for accessibility can also help us learn about effective ways of interacting with the mobile web. A general increase in the number and quality of voice interactive services will help to provide people with visual impairment better access to the web, as those services will have been designed for spoken interaction rather than being adapted from something designed for visual interaction. Elderly people and people with physical disabilities, who find access to the web via a keyboard and mouse difficult or impossible, and who struggle even more with small keypads on mobile devices, can benefit from multimodal interaction. Spoken or multimodal user interfaces also stand to benefit people with a low level of literacy. Core technologies required for multimodal browsing can support access to communication and web services by people who are hard of hearing - for example, an animated synthetic face alongside speech enhances intelligibility. They also support non-speaking people, who need speech synthesisers with prosody and personality, as well as effective user interfaces for data entry.

One way in which HP is addressing these issues is the Voice Web Initiative, part of HP's Philanthropic program, http://webcenter.hp.com/grants/. HP Philanthropy Initiatives focus on developing and supporting programs and partnerships that promote educational opportunity and e-Inclusion for people in underserved communities. The Voice Web Initiative aims to help open the Internet to a much wider section of society using voice and multimodal technologies and to help to develop a European research community and education programs in relevant technical areas. It has helped to support eight projects in 2001, at European universities and non-profit organisations, with a similar number in 2002. Many of these projects are researching or developing technology or applications for the elderly or people with disabilities, including: a synthesized talking head to aid intelligibility of telephone speech for people with hearing impairment; a voice driven Slovenian text-to-speech system for blind or visually impaired people; speech signal processing to enhance conventional hearing aids; voice interface design to help severely disabled users access the Internet; accessible school e-books for young students with visual disabilities; multimodal navigation for elderly people; a web based reading service for people with reading difficulties; web-based multimedia communication for non-speaking people.

## *Evolution – how will we get to the multimodal web?*

In order for multimodal interfaces to become widespread, we need devices and browsers to cooperate and new browsers to evolve, to provide a uniform service experience regardless of the interface modality, and we need standard ways of authoring these multimodal services. A number of factors will help to pave the way for an increase in interactive multimodal services. There are several relevant standardization or industry initiatives underway in: voice service authoring (mark-up languages for dialog, grammars and speech synthesis); multimodal service authoring (markup specifications for synchronization across multiple modalities); distribution of the speech recognition process between the end user device and the network; protocols for control of speech recognition and text-to-speech synthesis as distributed network resources. Speech and language technologies are developing incrementally, including recognition and synthesis, natural language understanding and generation and dialog management. This will increasingly move the quality and robustness of voice services from awkward, system directed dialogue, where the system asks all the questions, to more natural dialogues where a user can say what they want, when they want to, initially within the bounds of specific and limited application domains. Other recognition technologies, such as gesture, are also developing. Richer multimodal interfaces present challenges in fusion of information from multiple input sources and media planning and synthesis for system responses across multiple modalities.

One possible evolution is that, for most services, the coupling between components of a multimodal user interface would be weak at first, with different modalities being used to access a service at different times. Later, modalities would be more tightly coupled. Progress is being made - for example, at HP Labs, we've built early experimental prototypes that demonstrate multimodal service architectures and authoring, with voice and graphical browsers co-operating and ad-hoc connection and disconnection of devices.