Speech-enabled e-Commerce for Disabled and Elderly Persons

Georgios Kouroupetroglou and Evangelos Mitsopoulos
Universiy of Athens, Department of Informatics,
Panepistimioupolis, Ilisia, GR 15784, Athens, Greece
Fax: +30 1 6018677, E-mail: koupe@di.uoa.gr

1. Introduction

The term *e-commerce*, as used in this paper, implies a wide context of procedures performed electronically including:

- *Business-to-business* and *business-to-consumer* exchange of tangible or intangible goods.
- Financial services, such as *e-banking*, *e-investment*, or *e-insurance*.
- Information services, like *e-publishing* (newspapers, magazines, e-books etc).
- Entertainment services, such as *games*, *e-travelling* (electronic reservation of tickets, accommodation etc).
- Personal services, like *e-education, e-learning* and *e-training*.
- Miscellaneous services, such as *e-community*, *e-government* etc.

On the other hand, some authors refer to e-commerce using the term *e-business*, narrowing the meaning of e-commerce down to *e-retailing* [Cunningham et al 1999]. Indisputably, the exponential growth of e-commerce implies that in the short term it will become widespread and readily available to everybody. However, many of the exciting possibilities promised by e-commerce may be denied to a significant number of people, due to the substantive lack of appropriate consideration of their needs by the e-commerce researchers and developers [Noonan 1999].

A substantial proportion of the European population belongs to the disabled or the elderly user group [Roe et al 1995]. Moreover, there is a link between ageing and disability. E-commerce for the disabled and the elderly has much to offer to them, such as:

- Shopping, including selecting goods, accessing catalogues, paying for goods, home delivery options, etc.
- Banking and Finance, including selecting a bank, ATM issues, telephone and Internet banking, access to brochures and statements.
- Government information and transactions.
- Participation in employment.
- Implications of electronic publishing; and
- Access to other applications and services based on emerging technologies, including Java, operating systems for handheld devices, information kiosks, screen and web phones, XML, etc.

What differentiates the disabled and elderly groups from ordinary users is the low acuity or even the lack of some channels of communication, such as vision or hearing, or restricted mobility. This can be merely a result of ageing, or due to an accident, or perhaps a congenital condition. For the elderly and the disabled, it is extremely interesting to examine e-commerce from an accessibility perspective. Speech Technologies may offer a number of advantages and

access alternatives for the disabled and the elderly [Kouroupetroglou, 1995]. In this paper, the possibility of enabling access to e-commerce using speech technologies is investigated.

Some users are mobility impaired with reduced function of legs and feet or hands and arms (that is reduced dexterity). This condition can be congenital, due to an accident, or a matter of ageing. For instance, arthritis could result to a low dexterity or, in severe cases, lack of mobility. As far as communication channels are concerned, mobility impairment makes difficult or impossible to use devices, such as the keyboard or the mouse, commonly found in typical configurations for e-commerce. Visual impairments may range from partial to total loss of vision. Ageing typically leads to low visual acuity, too. In the case of total loss of vision, the visual channel, which is typically used by today's e-commerce applications to convey information, has to be substituted. In the case of low vision, magnifiers and enlarged text are not always convenient especially in the case of mobile devices the displays of which are of restricted size. Hearing impairments may range from hard-of-hearing conditions to deafness (profound loss of hearing). In the later, speech and sounds are inappropriate for presenting information. However, for many mild hearing impairments, including those imposed by ageing, spoken presentation of information does not pose any further problems than the use of a common telephone device. Regarding speech impairments, those associated with motor-speech disorders rather than a higher language or cognitive level disability, could take advantage of recent achievements in speech recognition [Hawley 1999].

The low acuity or the lack of channels for communication, like the visual or the motor channel potentially, renders speech technologies more appropriate than the visuo-motor communication required in standard desktop platforms for many elderly and disabled user groups. Indisputably, access can be attained in many other ways, too. However, speech technologies are more attractive since they would allow for universal design and access by most users groups regardless of any disability or their age [Lucente, 2000]. To this respect, this paper focuses on speech technologies in e-commerce. First, e-commerce platforms and services are briefly presented. In the next section, recent advances in speech technologies are examined. In the following section speech-enabled e-commerce products, protocols and standards, as well as services are critically discussed.

2. E-commerce

In this section, the various e-commerce platforms are briefly presented and along with the possible e-commerce services.

2.1 E-commerce platforms

There are a number of platforms on which e-commerce applications could be developed, including:

- common terminals,
- telephone terminals,
- information kiosks.

These platforms can be properly adapted for special user groups, such as the elderly or the disabled. From this point of view, this paper deals with a particular type of adaptation, namely the use of speech technologies for input and output. Indisputably, other adaptation technologies exist, but speech could additionally provide a uniform access for most user-groups (including the elderly and many types of disability) rather than merely an adaptation. Hence, this paper

investigates the application of speech technologies only. Furthermore, regarding the case of telephone terminals, of particular interest is the use of mobile telephone devices in e-commerce, namely mobile commerce.

A **common terminal** is the platform that has mostly been used for e-commerce applications, at least until recently. This platform offers a rich graphical environment which allows for proper presentation of graphical information, such as maps, graphs and pictures which have been developed and optimised for the visual channel of communication [Tufte 1997]. Apparently, where the visual channel of information is unavailable or visual acuity declines, alternative types of presentations need to be developed. Cluttered displays are very difficult to handle anyway. Using additional channels of communication, such as speech or sounds, could alleviate this problem. As far as the input devices are concerned, the keyboard and mouse are the typical ones. However, this is not necessarily the case with a number of disabilities where the mouse or the keyboard is completely inappropriate and different modes of input are required, one of which could be speech. Hence, it would be fair to say that speech input and output appear appealing in the case of common terminals for most users (but apparently not all, e.g. the deaf users).

Telephone terminals offer many advantages over common terminals when it comes to accessing e-commerce applications from any place, any time. Using speech technologies for input and output, e-commerce applications could require nothing more than a simple telephone device. On top of the existing network of fixed-point telephone devices, there is a continuously growing number of mobile telephone users. Importantly, the number of mobile phones already outnumbers that of common terminals by 4 to 1 [Boothroyd 1999]. In other words, e-commerce over telephone terminals may reach a much wider audience than common terminals. Moreover, telephones can be found almost anywhere; everyone could carry their mobile phones with them all the time. Common terminals, even in the case of laptop computers, are not easy to access from anywhere. For instance, checking for an e-mail message while sitting in a hotel lounge is much more conveniently done using a mobile phone that is more discrete. Due to their aforementioned advantages, telephone terminals constitute a particularly interesting platform for e-commerce applications and mobile commerce is a field of rigorous research. As far as speech technologies are concerned, they have plenty to offer to mobile terminals. The later have particular limitations both in their input (restricted keyboard which renders text input very difficult and time consuming) and output (limited size and resolution of their visual displays). Many of these problems could be alleviated by the use of speech technologies.

Information kiosks are not without problems, either. Their virtual on-screen keyboard is not easy to use effectively for many users. Also, the public behaviour of users is radically different to that in private. They are more hesitant to experiment with the system in order to avoid errors, since they don't want to appear less knowledgeable or capable than the other ones. Speech may have to offer more intuitive dialog interfaces for information kiosks and for telephone terminals.

2.2 E-Commerce Services

Regardless of the particular type of an e-commerce application, a number of *e-commerce services* are likely to be brought together under that application. In other words, an e-commerce application can be decomposed down to its ingredient-services. The most important services include:

- E-catalogs,
- E-forms,
- E-mail,
- Individualisation of services including:
  - User authentication,
  - User registration,
  - User profile,
- Search,
- E-basket,
- Check-out (finalise order and pay for the goods in the e-basket).

In particular, e-forms are the building blocks of all the services above apart from e-catalogs since, all these services rely on or can be reduced to an e-form, at least as a first approximation [Ström 1998].

E-catalogs: E-catalogs are an indispensable feature of an e-commerce application. They provide detailed information for the products on offer, their shape, content and price, as well as the option for ordering and paying for them on-line. It allows their customers to investigate and compare products at their own convenience without having to confront a salesperson or to physically be at the store. A smart e-catalog takes advantage of many services including e-forms, e-mail, individualization, search, e-basket etc. to provide a flexible tool for identifying a proper product, notification of offers etc.

E-forms: E-forms are used when the user is required to provide data to the e-commerce application. They consist of a number of fields in which the user assigns values. These fields may have different forms including radio buttons, check boxes, text-edit boxes, free-text boxes, buttons, etc.

E-mail: Competition in e-commerce forces companies to have a direct contact with their customers, and thus a better idea of their needs and preferences, in order to keep them *loyal*

to the company and away from the competitors. E-mail offers the opportunity of such a direct contact. It is a combination of e-forms and free-text boxes.

Individualization: In many cases it is necessary to perform user authentication, that is to prove that the user is the one he or she claims to be. One obvious example is payments or access to members' areas only. Currently, the typical way of authentication is the use of a username and a password. However, speech, among other biometric technologies, offers an alternative solution. Some e-commerce applications invite the user to register with them. In this way, they can obtain a user profile of the preferences, needs and requirements of each particular user, and tailor the information presented to them accordingly. Individualization relies on completion of e-forms.

Search mechanisms: An essential feature of e-catalogs that facilitates and accelerates the finding of the required information or product. For instance, in an electronic bookstore, one may search for books of a particular theme, of the same author, as well as for a book using its ISBN, keywords etc. Again, these rely on e-forms.

E-basket: Instead of repeating the payment process for each product they buy, customers use e-baskets. E-baskets are associated with certain functionality for adding and removing products as well as examining those in the basket. They are based on e-forms.

Payment: When the user decides to purchase the content in the e-basket, the detailed charge including taxes and delivery cost are presented. Usually, there is a choice for several credit cards and sometime for digital money. In any case, an e-form has to be completed.

3. Speech Technologies

The term "speech technologies" involves mainly two different technologies associated with speech, namely speech recognition and speech production. In the case of speech recognition, speech constitutes the input to the system, whereas speech production provides a modality for conveying the output of the system verbally to the user through the auditory channel. Apparently, both technologies can be integrated in a system to form a uniform – in terms of modality – channel of communication or accessibility with the machine, named: speech interfaces, natural spoken dialogue systems or conversational interfaces [Lai, 2000], [Boyce, 2000].

3.1 Speech recognition

Generally speaking, the objective of a speech recognition system is to recognize some aspects of a spoken utterance. In practice, speech recognition may have different applications such as:

- **Dictation**, where the system transcribes the utterances spoken by the user into text,
- **Speech understanding**, where the system attempts to understand the meaning of the utterance spoken rather than to transcribe it,
- **User recognition**, where the system attempts to recognize the user from his/her *voiceprint* and, finally,
- **Language recognition**, in the case of multilingual systems.

Regardless of its application, a speech recognition system can be characterized along a number of dimensions that influence the precision with which the speech recognition task is performed and characterize its appropriateness to a particular application. The most important parameters include:

- **Speaker dependence / independence**, that is, whether the system has to be trained for each individual using it, or it may recognize different speakers without having a different model for each individual one. Whether or not speaker independence is a desirable feature or a requirement of a speech recognition system depends on the context of the interaction taking place. Clearly, a public access system cannot be speaker dependent, whereas most dictation software is actually trained on a user basis. Although, speaker independence offers the significant advantage of not requiring separate training for each user, it is less precise in order to allow for inter-speaker variations. An important consequence of this is the limited

  vocabulary size a speaker-independent system can support in comparison with a speaker-dependent one.
- **Vocabulary size**. Whether a small or a large vocabulary is required depends on the application. In a dictation system, it would be desirable to have a size of tens or of hundreds thousands words. However, the larger the vocabulary the system supports, the lower its accuracy and recognition speed become.
- **Discrete / continuous speech**. If words in a utterance have to be spoken isolated from each other in order to be recognized, then the speech recognition system offers discrete speech recognition. On the other hand, if the user can articulate the utterance in a natural way without having to stop in between words, then speech recognition is said to be continuous. Discrete speech recognition is simpler than continuous and, until recently, significantly more precise, too. However, relatively recent techniques such as word spotting and sub-word modelling [Markowitz 1996] have rendered continuous speech recognition particularly flexible and even dictation systems (which support large vocabularies) allow for continuous speech recognition.
- **The quality of the communication channel**. For instance telephone speech today still offers limited bandwidth. More importantly, the level of the background noise is important since the higher it is the more probable a miss-recognition becomes. Recent developments allow for robust recognition systems that can resist even to high-level background noises.

Given these dimensions, it is important to see the limitations posed by the state of art technology on the aforementioned types of speech recognition applications.

3.1.1 Dictation

In dictation applications, the system transcribes what the user dictates to it recognizing each word being spoken to it. There are many situations in which speaking an utterance is deemed to be more convenient to the user than typing it using a keyboard. In some cases, dictation is faster than typing or ergonomically better. Sometimes, a proper keyboard cannot be incorporated in the system – many people have experienced the difficulty of writing messages using the numeric keypads of their mobile phones.

In any case, a fundamental requirement of dictation applications is a large vocabulary while preserving high recognition accuracy. A second fundamental requirement is continuous speech recognition, to increase the naturalness of the interaction and, consequently, to allow the user to focus on the content of the utterances rather than on the manner they are uttered. However, in

many cases, speaker independence, although desirable, is not a hard requirement (apart from the public access scenario). In fact, it is a current technological limitation, that large vocabulary

speaker independent dictation has yet to be achieved. Nevertheless, dictation systems have become quite flexible on this respect, since they require a minimum amount of initial training and they keep on improving their user models and hence their precision during actual dictation tasks. A number of commercial state-of-art dictation systems is presented in Table 1.

System	Naturally Speaking	Viavoice	VoiceXpress	FreeSpeech 2000
Vocabulary size	250,000	260,000	220,000	300,000
Speech type	Continuous	Continuous	Continuous	Continuous
Recognition accuracy	98%	94%	98%	n/a
Initial training	5 min	n/a	12 min	15 min
Languages supported	English German French Spanish Italian Dutch Australian Japanese Swedish	English German French Spanish Italian Chinese Japanese	English German French Spanish Dutch Chinese	English German French Spanish Italian Dutch Austrian Belgian Swedish Portuguese Brazilian

Table 1: Comparison of four commercial dictation systems. Note that there are no substantial differences between them (those appearing in this table are mostly due to the different sources referred to [H1], [H2], [H3], [H4])

Summarizing the state of art regarding commercial dictation applications, current systems offer large vocabularies (≥ 200,000 words) and continuous speech recognition in a variety of languages. At a research level, dynamic programming methods [Deshmukh 1999] may allow for large vocabulary speaker-independent dictation. These techniques are likely to be adopted in commercial products in the near feature.

3.1.2 Speech Understanding

In contrast to dictation, speech understanding aims at extracting the *meaning* of the speaker's utterance, rather than identifying every single word spoken. As a consequence, in this case, speech recognition technologies have to be integrated with natural language processing technologies so that the speaker's utterances can be identified on the basis of a few keywords, as well as the context emerging from the dialogue discourse that have been carried out so far with the system. Hence, speech understanding involves access to a number of different sources of information including the lexical, syntactic and semantic level of the utterance, and, preferably, its pragmatic context and the prosodic information conveyed with that utterance. An example of speech understanding provided by the JUPITER research system is presented in Table 2.

Speaker	Yes, I would like the weather forecast for London England, please.
System	In London, in England, Wednesday. Partly cloudy skies with periods of sunshine. High 82 and low 63. Is there something else?
Speaker	What is that in degrees Celsius, please?
System	In London, in England Wednesday. High 28 Celsius and low 17 Celsius. What else?

Table 2: Speech understanding in JUPITER [Zue 2000]

JUPITER has a vocabulary of approximately 2,000 words. Since May 1997, it has received more than 100,000 calls achieving accuracy of 89% in recognizing words (which rises to 98% for trained users). Regarding the successful recognition of the meaning of utterances, accuracy percentages are 80% and 95% for novice and experienced users respectively. It should be noticed that recognition percentages are lower for non-native speakers, though.

Speech understanding has been put into use in other contexts, too. One instance is the Workshop on Automatic Speech Recognition and Understanding (ASRU'99) [H5] where registration and information requests about accommodation, travelling and social events related to the workshop were handled by a speech recognition system with speech understanding capabilities. The system also allowed the authors to be informed for the status of their submitted papers and some further functionality regarding the registration of members of IEEE.

3.1.3 User Recognition

User recognition refers to the process of relating a particular person with its true identity. There are two different cases of recognition: speaker verification (authentication) and speaker recognition (identification). In the case of speaker verification, the system has to answer the following question: "is this user the one he/she claims to be?". On the other hand, speaker identification attempts to successfully answer the question: "who is this speaker".

Regarding e-commerce, the most frequent scenario refers to speaker verification (authentication). Traditionally, this can be accomplished in two different ways, either using some object like and ID card or a passport, or some knowledge which only the system and the

user posses (for instance, a password). However, in the last few years, there have been substantial improvements in a number of biometric identification techniques [Markovitz, 2000], which rely on the physical and behavioural characteristics of a person that are deemed to be unique to that person.

One important biometric identification technique relies on the *voiceprint* of the user. However, this technique has some limitations because the natural features of voice vary with background noise, the type of the microphone used, the health condition of the vocal tract of the user etc. As a consequence, a system has to allow for some variations in the user's voiceprint. But as permissible variations become larger, the more the possibility of some other user being miss-authenticated by the system and hence less the security of user recognition offered by the system. Thus, current systems frequently combine biometric with traditional authentication to increase their security.

A user-recognition example is provided by Philips (SpeechWave software package [H6]). The first time the users call the speech recognition system, they have to register with it. To do so, they have to repeat they password a number of times in order for a reliable voiceprint to be acquired by the system. This is illustrated in Table 3.

System	To enrol to the voice banking system, we first have to take your voice print. Please say your six-digit voice banking ID number now.
Speaker	One two three four five six.
System	Please, repeat your voice banking ID number now.
Speaker	One two three four five six.
System	Your voice print has been taken. Please say your voice bank ID number now.
Speaker	One two three four five six.
System	You are admitted to the system. How may I help you today?

Table 3: User recognition by the SpeechWave system [H6]

Nuance [H7] offers the following interfaces in its product "Nuance Verifier" [H8] as shown in Table 4.

Interface	Registration	Recognition
User defined password	Speak password 2-3 times	Speak password 1-2 times
User name	Speak your name 2-3 times	Speak your name 1-2 times
Account number + random digits	Speak account number plus 3-5 sets of random digits	Speak account number plus 1-2 sets of random digits
Random phrases	Speak several sentences of text	Speak 2-3 sets of phrases

Table 4: Registration and recognition methods offered by Nuance Verifier.

A number of commercial products exist ranging from inexpensive simple systems like voicecrypt 2.01 of Veritel [H10] which may offer some protection to files but it is far from being secure. On the other hand, there are large and expensive systems like Citadel GateKeeper of InteliTrak Technologies Inc. [H11] which is one of the most secure systems tested by PC Magazine [H9] and may cover the needs of up to 5,000 users. However it is much more expensive – a Citadel would cost at least as much as 1,000 voicecrypts.

3.1.4 Language Recognition

In this case, the objective is to infer the language spoken by the user, on the basis of a piece of text spoken by that user. There have been significant advances, at least on the research frontier. For instance the system PRLM-P [Zissman 1997] offered recognition accuracy of 74.3% for a 30 sec utterance and 53.4% for a 10 sec utterance, for 12 languages. Its accuracy in discriminating between two languages is more than 90%. Moreover, a different system supporting 4 languages achieved a 90% correct recognition rate for 10 sec utterances [Corredor-Ardoy 1997]. As far as commercial systems are concerned, the language recognition system by Sanders [H12] may distinguish between 11 languages with accuracy 85.4% (10 sec utterances) and 93.3% (50 sec utterances).

3.1.5 Human Factors in Speech Recognition

These can be classified according to [Markowitz 1996]:

- The capabilities and limitations of the underlying speech technologies. Most requirements can be effectively fulfilled by state-of-art speech recognition; currently, there are significant efforts expended to render dictation speaker-independent.
- The appropriateness of speech technologies relevant to the tasks supported, including the type of information to be presented, navigation in the user interface, and error recovery; and
- The needs and preferences of users. Different user groups may have different needs to cater for. Moreover, there should be a distinction between novice users which require more help to perform a task, and expert or power users which could take advantage of interaction shortcuts, etc.

3.2 Speech Production

Speech production can be based on three different methods:

- Waveform coding
- Analysis-synthesis
- Synthesis-by-rule

A comparison between these methods is shown in Table 5.

Features	Waveform coding	Analysis-synthesis	Synthesis-by-rule
Quality {understandability {naturalness	 High High	 High Medium	 High Medium
Vocabulary size	<500	1000s	Unlimited
Bit rate	24-64kbps	2.4-9.6kbps	50-75bps
Duration stored in 1Mbit	15-40 sec	10sec – 7 min	Unlimited
Stored units	Syllables, words, sentences	Syllables, words, sentences	Phonemes, syllables etc.
Complexity	Low	Medium	High
Hardware	Memory	Memory and CPU	CPU

Table 5: Some Features of Speech Production Methods

3.2.1 Waveform Coding Method

The initial step in this method is the recording and digital storage of appropriate units of human voice like words or whole phrases. Following that, the desirable utterances can be spoken by locating the waveforms of the appropriate units and concatenating them. A fundamental factor in waveform coding is the size of the units that will be recorded, which may range from phonemes to whole phrases. The size of these units significantly affects the quality of the speech produced as well as the flexibility of the system in composing different utterances. If whole phrases were used, the quality of the speech would be particularly high, but the number of different phrases the system would be able to produce would be severely limited. The most flexibility could be achieved by using phonemes as units but the quality of speech produced would decay substantially.

Trading off flexibility with quality implies using medium size units like words. However, if speech is to appear natural as well as understandable special attention is required during the recording phase for a number of reasons. First, uttering a phrase as such sounds different than uttering it word by word. Words in a utterance have almost half the duration than when spoken isolated, which results to waveform coding with words as units to sound painfully slow. Second, even if different variations of the same words have been recorded, it is still difficult to achieve the appropriate intonation and prosodic features the syntax and the semantics of a phrase would require.
For these reasons, commercial systems (for instance SpeechWorks [H13] and others) typically combine small phrases with words. Until recently, the reason for choosing waveform coding over other methods was the good quality of speech offered. Yet, a waveform coding system may have a limited vocabulary and cannot convert any piece of text into speech. For instance, this method would be appropriate for a menu-based application presented over the telephone but inappropriate for a screen-reader application. Moreover, advances in synthesis-by-rule methods has resulted in improved speech quality and rendered these systems attractive for commercial applications, too.

3.2.2 Analysis-Synthesis Method

In this case, pre-recorded speech is analysed and converted into sequences of parameters. The speech synthesizer is driven by these parameters, properly placed into sequences according to the desirable message to be produced. Since waveforms are converted into sequences of

parameters, the amount of information that has to be stored is much smaller than that required by the waveform coding approach. Nevertheless, the quality of speech decays.

3.2.3 Synthesis-by-Rule Method

In this method, the most important parameters for the fundamental units of speech such as syllables or phonemes are stored, and during speech reproduction, they are concatenated by use of a number of rules. These rules are further modified as well as influenced by proximal units to support intonation and prosodic variations. In theory, this synthesis method could rely on phonemes only, the number of which is limited for each language – usually less than 50. However, the phoneme concatenation rules are very complex and not well understood, which leads to low quality speech. For this reason, diphones are mostly used, which results to understandable speech of not particularly high naturalness. Nevertheless, a number of speech synthesizers are available including those by AT&T [H14], Bell Labs [H15], Eloquence [H16], Apple [H17], to mention just a few. Moreover, screen-readers like JAWS [H18] and HAL [H19] have been using speech synthesis.

A recent development, the unit selection method, allows for different unit sizes to be used. The sizes can range from phonemes to whole phrases and are selected appropriately by the speech synthesis system depending on the intonation and the desirable prosodic features in each case, so that a minimum of modifications are required during concatenation. As a result, this method results to high quality speech both in terms of understandability and naturalness. At the present, a drawback of this method is that creating a voice is an expensive and time-consuming process. Moreover, the process has to be repeated for every single voice the synthesizer supports. For instance, for the RealSpeak synthesizer [H20] by Lernout & Hauspie [H21] a voice requires about 2 months of recording sessions, resulting in approximately 2Mb of speech samples that will be used by the synthesis algorithm. Nevertheless, as synthesis-by-rule attains more and more natural speech, it tends to become more suitable than waveform synthesis for e-commerce applications, since it is flexible for dynamic content.

3.2.4 Human Factors in Speech Production

Speech is not appropriate for all types of information. For instance, a photograph can only be described by speech. The same holds true for diagrams, maps etc. depending on the type of information the user intends to extract from these representations. Sometimes, spoken presentation is time consuming – the designer could opt for merging speech with sounds (which convey information) to improve the aural presentation [Mitsopoulos 2000]. The differences between static visual displays the content of which exists over time and dynamic aural presentation which evolves over time may imply that some visual interaction techniques are inappropriate for spoken presentation of information. Finally, multimodality and its potential benefits is still a research issue (incorporated in the W3C agenda) regarding both speech input and output modalities used in conjunction with other channels of communication like vision and touch.

4. Speech-enabled E-commerce

Many corporations actively involved in speech technologies taking into account the forecasts for the exponential growth of e-commerce business have already envisioned the application of these technologies. As a consequence, they have taken steps forward to form strong co-alliances and develop a number of standards and protocols.

4.1 Speech-enabled e-commerce products

Intel has invested 30 billion dollars on Lernout & Hauspie Speech Products to develop a multilingual environment for placing orders [H22]. Vocalis, which specializes in speech recognition, has developed SpeecHTML to allow internet access using speech input and output. In 1999, over 30 ISPs in the U.K. provided telephone access to the internet using SeecHTML [Boothroyd 1999]. Recently, two well known portals (Lycos and Yahoo) have launced voice portals, too. Using its voice recognition software (ViaVoice), IBM has developed a number of on-line demonstrations of speech-enabled e-commerce applications including directory dialling and mutual funds [H23], as well as a number of CRM applications [H24] such as CallPath, DirectTalk, Directory Dialler, Mail Analyzer etc. CrossMedia Networks Corp. is developing applications like MyInBox Voice Email [H25] that allow access to the internet and e-mail over mobile telephones using speech technologies for input and output.

Regarding the research frontier, the Spoken Languages Systems Group at MIT University [H26], has developed speech dialogue systems like GALAXY which allow the user to interactively access information. Built on GALAXY are a number of demonstrations such as JUPITER (weather forecasts), PEGASUS (flight information), DINEX (restaurants in Boston), WHEELS (small ads for cars) etc.

4.2 Speech-enabled e-commerce protocols and standards

In order to achieve speech-enabled e-commerce and in particular mobile commerce, there are a number of protocols and standards from different areas that need to be brought together. In general, speech technologies applied on e-commerce would require the implementation of some form of voice browser. Hence, it is interesting to see whether suitable protocols and standards have been developed as far as the incorporation of speech technologies in e-commerce is concerned. Regarding m-commerce, wireless protocols become important, too. Moreover, the convergence of voice protocols and wireless protocols are of prime importance, that is how a mobile device despite its limitations, could support speech technologies, as well as what is expected in the foreseeable feature.

4.2.1 Speech technologies related protocols

There are a number of protocols, most of which form extensions of XML, which is regarded as the standard for e-commerce applications. Here, SAPI of Microsoft, JSML by Sun Microsystems, voxML by Motorola, and VoiceXML by the voiceXML forum will be mentioned.

- **Microsoft Speech API (SAPI)**
  SAPI [H27] is provided by Microsoft for the Windows operating systems for applications that employ speech technologies. It consists of the following blocks: Voice Command, Voice Dictation, Voice Text, Voice Telephony, DirectSpeechRecognition, DirectTextToSpeech, and AudioObjects. The SAPI architecture is shown in Figure 1.

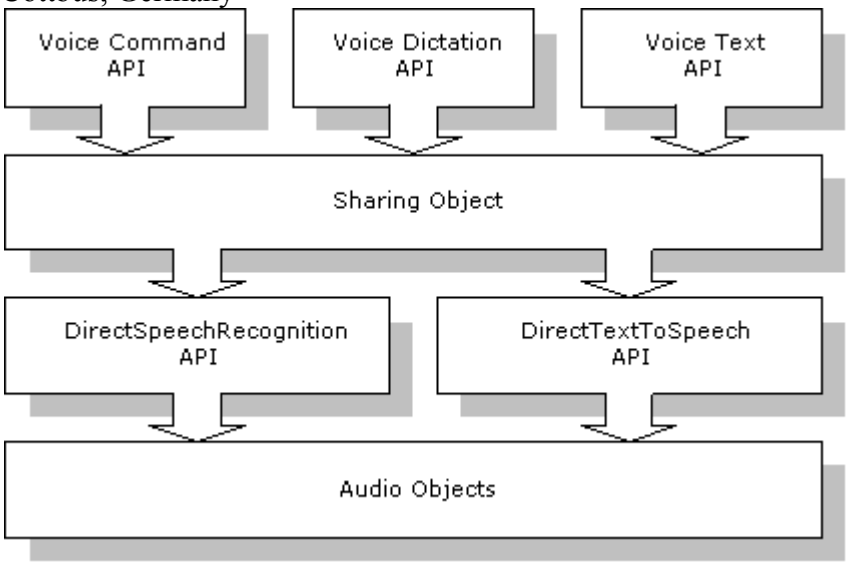

Figure 1: Microsoft Speech API

- **Java Speech Markup Language (JSML)**
  Sun Microsystems has developed the JSML protocol [H28] in 1997 which incorporates information about the pronunciation of words, prosodic cues etc. It is an XML extension, which renders JSML independent from the Java Speech API [H29]. Importantly JSML has been acknowledged by W3C which plans to employ it in its "Voice Browser" activity in the speech grammars part (see below).

- **VoxML by Motorola**
  VoxML [H30] extends XML for both speech recognition and speech output. In Figure 2, the architecture proposed by Motorola in [H31] is shown.

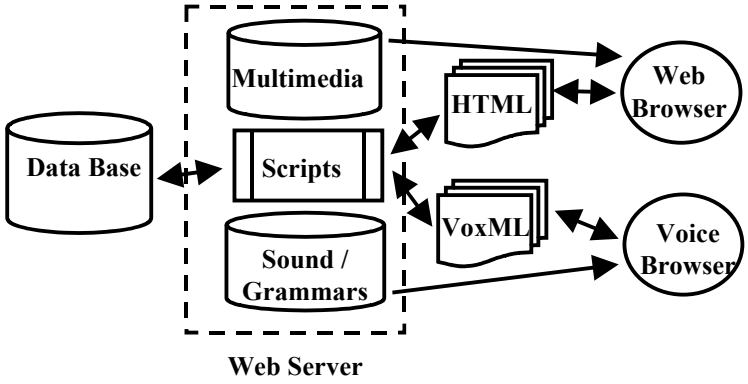

Figure 2: voxML and Motorola's architecture.

- **VoiceXML**
  In March 2000, the voiceXML forum announced v1.0 of voiceXML which is a result of collaboration between 79 participants including AT&T, Motorola, IBM, Lucent Technologies and others [H32]. Importantly, voiceXML has been acknowledged by W3C which plans to adopt it as a dialog markup language for interactive voice-response applications (in the W3C "Voice Browser" activity) [Lucas, 2000].

In general, there are a plethora of issues related to Voice Browsers. The relevant W3C group working on the "Voice Browser" activity [H33] has released a collection of requirement drafts regarding voice dialogs, reusable dialog components, speech grammars, natural language representation, speech synthesis, multimodal systems, and voice browser architectures.

It is envisaged that the work by the W3C Voice Browser activity will be applied to m-commerce, too (for instance, see below for the collaboration between the WAP and W3C forums). In the experimental stage, there are also languages like TalkML [H34] by HP Labs for use in markets like call centres (IVR), smart phones with displays, in-car internet access systems, and mobile devices with restricted displays or keyboards. Regarding voice browser architectures and, in general, architecture for speech-enabled e-commerce, one should also refer to the white paper by the V-commerce alliance [H35] with participants such as Nuance, Motorola, NetSage, VISA International, Vocalis and others.

4.2.2 Mobile devices and wireless protocols

Several different wireless protocols are emerging; here the SIM Application Toolkit, the WAP, and the MExE are briefly presented.

- **SIM Application Toolkit**
  It has been agreed and incorporated within the Global System for Mobiles (GSM) standard. "SIM" denotes the smart card inserted into the GSM mobile phones that contains information about the user. SIM Application Toolkit allows the flexibility to update the SIM to alter the services and download new services over the air. However, significantly, two of the three largest mobile vendors, Ericsson and Nokia, have not launched or announced SIM Application Toolkit compliant phones [H36].

- **Wireless Application Protocol (WAP) by the WAP forum** [H37]
  Motorola, Nokia, Ericsson and the US company Phone.com teamed in 1997 to develop WAP. It is nowadays widely accepted and widely hyped in the mobile industry and outside of it. WAP is an attempt to define the standard for how content from the internet is filtered for mobile communications. It is simply a protocol – a standardized way that a mobile phone talks to a server installed in the mobile phone network. It takes into account the particular features and limitations of mobile devices like their small size, limited display, lack of mouse, small keyboard, limited CPU, memory and battery life, network instability etc. Its scripting language is WML (wireless markup language), which is compatible with XML. The adoption of XSL (extensible Style Language) will offer a mechanism for automatic conversion from well formed XML to WML. WAP does not de facto support speech technologies but it has the potential to do so. WAP and W3C are in close collaboration on this issue. In September 2000, a W3C/WAP Workshop ("the multimodal web") took place in Hong Kong to examine the current situation and establish a research agent of issues like the integration between WAP and speech technology languages like voiceXML H38. Hence, a significant effort is on its way to incorporate speech technologies in mobiles and make them truly multimodal in the foreseeable feature.

- **Mobile Station Application Execution Environment (MExE)** [H39]
  MExE is a wireless protocol that is designed to be incorporated into smart mobile phones. It builds a Java Virtual Machine into the client mobile phone. It supports a wide range of interaction methods including voice recognition, icons and softkeys. MExE and WAP share several similarities but have some fundamental differences. Whereas WAP incorporates some scripting, graphics, animation and text, MExE allows full application

programming. As a consequence, because of the significant processing resources required on the mobile client, MExE is primarily aimed at the next generation of powerful smart phones. Because the processing power to run Java applications is not currently available in mobile terminals, MExE will be fully utilized further in the future than WAP.

In conclusion, regarding the protocols and standards for speech-enabled e-commerce and mobile commerce, there has been a lot of effort on standardization aiming at the derivation of suitable, widely accepted, protocols for speech-enabled e-commerce and mobile commerce in the foreseeable feature.

4.3 Speech-enabled e-commerce services

Having examined the protocols and standards that could be used to incorporate speech technologies in e-commerce, we proceed to an issue relevant mostly to the human-computer interaction perspective. How these technologies could be appropriately applied to enhance the usability of each service? In the following sections, the e-catalog, e-form, e-mail and individualization (in particular user authentication) services have been selected out of all the services presented in Section ??. E-catalogs focus mostly on issues relevant to the verbal presentation of information. Nevertheless, user input using speech recognition technologies is also relevant and it is further examined in e-forms. As it has been already mentioned, e-forms can cover the rest of services, at least at a first approximation. However, the e-mail service is also presented due to its special requirements regarding speech recognition. For similar reasons, user authentication is investigated, too.

4.3.1 E-catalogues

E-catalogues are composite services comprising of other simpler ones such as e-forms, e-baskets, e-mail etc. Simpler services are examined in the following sections. As far as the application of speech technologies on e-catalogs is concerned, the following issues are examined:

- The quality of speech production.
- The suitability of speech for presenting information.
- The problem of navigating in the catalog and error recovery.

Telephone terminals
This platform is significantly more widespread than common terminals and offers the possibility of e-commerce from any place at any time (for instance mobile commerce). The visual displays of mobile devices are of modest size, which makes them inappropriate for presenting most types of information apart from a few words of text at a time, as well as undesirable for the elderly and disabled users with low visual acuity or no vision at all. Moreover, their keypad is difficult to use to enter text (for instance messages); many elderly or disabled users would find it difficult to use even for dialling numbers.

The quality of speech production for presenting products in an e-catalog has to be high regarding both its understandability and its naturalness; otherwise the spoken presentation will be incomprehensible or tiresome to the users. Either waveform coding or synthesis-by-rule methods (and in particular unit selection methods) could provide such quality. The later is more flexible and hence more suitable to e-catalogs in which content is dynamic and may change at any time (which eventually renders waveform coding impractical).

Given the high quality of speech attainable, the next question is whether speech is appropriate for presenting the information conveyed in e-catalogs. Apparently, some types of information are hard to present verbally (for instance maps – refer to the human factors in speech production section). Nevertheless, even in these cases, a verbal description is better than nothing (given the limited display size, speech could be potentially more advantageous than graphical presentation even for these types of information). A concrete example is a blind user, using an e-commerce application to shop at a food store.

Spoken presentations evolve over time, that is, they are serial and thus time consuming. Consequently, users should have control over the presentation of information and interrupt it at their will. Furthermore, navigating in a spoken e-catalog has to be supported. There are similarities as well as important differences between vision and speech. As a result, just reading aloud a visual e-catalog does not retain the usability of the visual e-catalog. Users should always know their whereabouts in the e-catalog, either by explicitly requesting this information ("where am I?"), or because the interface has been designed to provide constantly this information in the background (for instance by using sounds – see [Mitsopoulos 2000, Stevens 1996, Brewster 1995, Pitt 1998]).

Since users have to issue commands to navigate in the e-catalog, but, in many cases they are unwilling or unable to use the keypad, speech recognition technological issues become important. In this case, users have to issue commands to browse the catalog. The technology required is speaker independent continuous *speech understanding*, which would allow for any user to phrase the commands in a natural way (not having to remember the precise commands). Keyword spotting and barge-in techniques would be useful enhancements. Public systems like JUPITER (refer to the speech understanding section) imply that these requirements can be fulfilled by the state-of-art technology. Error correction issues are dealt with in the e-forms case.

Common terminals

Although common terminals offer a rich graphical environment as well as keyboards and mice for input, many disabled and elderly users may either prefer or have to restore to speech technologies. Low vision would make graphically rich e-catalogs with small fonts difficult to read. Descriptions of products and any text in general could be read to the user on demand. For blind users, all the visual information has to be transformed in speech and sounds. For users with reduced dexterity, speech recognition could replace the keyboard or the mouse.

The technologies required in the case of e-catalogs have already been discussed in the previous paragraphs. Two additional issues should be mentioned here. The possibility of personalizing a common terminal implies that speaker-dependent recognition is feasible, yielding even better recognition accuracy. Secondly, an e-commerce application will be one of the many applications used by these users. As far as these applications are concerned, if users are better off by employing speech technologies in interacting with them, then the same type of interaction has to be preserved with e-commerce.

Information kiosks

Information kiosks may offer reach graphical environments just like common terminals. Hence, as discussed above, there are many cases where the disabled or the elderly user could find the use of speech technologies beneficial. A particular issue is the virtual keyboard displayed on the touch screen of a kiosk. Even normal users find it difficult to use; for users with even mild dexterity problems and reduced visual acuity the situation is much worse. Hence, it is important

to incorporate speech recognition technologies, and in particular, speaker independent continuous speech understanding with particular robustness to noise.

4.3.2 E-forms

Two interaction issues prevail regarding e-forms. The first one is the assignment of a value to a field in the e-form. The second issue is relevant to the orientation of the user while interacting with an e-form. E-forms can be simple, consisting of one or two fields only. On the other hand, they can be really complex, with groups of fields, optional fields, dynamic fields which alter the appearance of the e-form etc. It is necessary for the user to be aware of the structure of the e-form. How speech technologies can be applied on particular platforms in the case of e-forms is examined below.

Telephone terminals
There are three main issues regarding filling-in a field of an e-form. The user has to understand or identify what the field to be completed stands for. For instance, the field "Surname". Secondly, to assign a value to that field and, thirdly, to get a feedback for that value assigned. The second issue is related to speech recognition technologies, whereas, the other two refer to speech production technologies. The later is similar to e-catalogs, in the sense that high quality speech is required, as well as flexibility due to the dynamic attribute of these fields. Hence, synthesis-by-rule and preferably unit selection techniques are the best candidate. It should be noted here that possibly, most of the time, the name of a field as well as the value assigned to it could be presented on the display of a mobile phone. However, even common visual problems of ageing make it difficult to read the display, and thus many elderly and disabled users would prefer spoken presentations which have the advantage of being short compared to those in e-catalogs. User groups that have to use speech recognition, might prefer for ergonomic reasons the spoken presentation so that they don't have to switch between speech and the visual modality. In any case, speech can be presented simultaneously with text, to accommodate different user preferences.

As far as speech recognition is concerned, supplying the value to a field would most of the time rely on dictation since the precise value should be entered. In many cases, there are just a few values (that is a small vocabulary) that would allow enough robustness for a speaker independent dictation system of continuous speech. In some cases though, it is necessary to provide a large vocabulary – one case is the composition of e-mails which will be discussed in the following section. Moreover, it is important to find a mechanism for distinguishing whether the user provides a value to a field and whether his or her utterance is a command regarding the navigation in the e-form.

Yet another requirement for speech recognition, is the ability of the user to issue navigation commands at any time, to check the contents of a field or to move among them in the structure of the e-form. Similar considerations to those of e-catalogs apply. A continuous speech understanding speaker independent technology would be preferable. Combining these requirements together, one could say that most of the time the speaker independent continuous speech understanding technology would be appropriate to e-forms, but in some cases where dictation is required, this technology is still speaker dependent. More details are presented in the section on the e-mail service.

Considering the restricted input devices of mobile phones, speech can be a real boon when it comes to writing text rather than just simple numbers – this becomes even more dramatic in the composition of messages (and in general in filling-in a text box). This advantage is even more

pronounced in the case of reduced dexterity or visual acuity (both are frequent impairments in the elderly and the disabled user groups). Hence, speech input is of fundamental importance to normal users, and even more so to the disabled and the elderly user groups.

Finally, regarding the structure of an e-form, it can be dynamic and complex, just like the structure of an e-catalog. Along with the considerations presented in the e-catalog section, one could further notice that filling an e-form could be more system pre-emptive, that is driven by the system rather than the user.

Common terminals
The improvements introduced by speech technologies have already been discussed in the relative section on e-catalogs. As a general observation, the combination of speech technologies and a visuo-motor environment could help to accommodate many widely different user preferences and needs under the same e-commerce application. In other words, it is a mater of producing a more universal design [Stephanidis 2000].

Information kiosks
As with e-catalogs, virtual keyboards may pose a series of problems in entering text and values; speech technologies could help the disabled and elderly groups in a similar way.

4.3.3 E-mail

Despite that e-mail can be reduced to a relatively simple e-form, it is mentioned separately here for the reason that dictation speech recognition technology is required. Due to the extended vocabulary that might be used, it is not yet feasible to provide a speaker independent technology, which has different impacts on the three platforms considered here. One could also notice that, along with navigation commands, further ones are required to manage the mailboxes and the folders in which mails are kept. However, this does not further affect the speech technology required.

Telephone terminals
The fact that e-mails potentially constitute a large piece of text (more than a few words) has two consequences in its presentation and its composition. The restricted display of this platform allows for only a few words to be presented at a time; the user has to scroll, which implies further commands issued with the keypad. This can be a disadvantage whenever users have reduced dexterity. Partial presentation can also be annoying to some users, which may prefer an uninterrupted spoken presentation.

Accessories like small QWERTY keyboards attached to mobile phones simply demonstrate the problem that normal users face when it comes to composing an e-mail (and any message in general). Input of large texts is even more problematic for the user groups in focus. For them, it is fundamental to use speech recognition rather than the keypad.

Some steps have been taken towards voice browsers as has been reported in the speech-enabled e-commerce products sections. These allow for speaker independent access to e-mails, but at the moment composition of e-mails is not supported. This is because dictation still requires training. In public access systems this is not an option; currently research efforts are focused on eliminating the need for training but a commercial product has still to become available. On the other hand, mobile phones incorporating speech recognition in the client, could be trained by the user to perform all speech recognition tasks on the mobile phone rather than on a server.

Common terminals
In many cases, common terminals are not used publicly – or a user profile can be stored. Thus, dictation *can* be speaker dependent, and e-mail can be given full functionality. In general the observations on e-forms and e-catalogs apply in this case, too.

Information kiosks
As with telephone terminals, virtual keypads could be substituted with speech recognition technologies.

4.3.4 Individualization

Although individualization of services is a composite issue, this section focuses on user authentication and, in particular, on the use of speech biometric technologies to substitute the traditional system of usernames and passwords. Having to learn and remember passwords can be difficult for the elderly (and not only them!). Moreover, biometric technologies offer, in principle, increased security. Furthermore, the need for typing a password (which can be particularly difficult with the restricted keypads found in telephone terminals or for people with reduced dexterity) is eliminated.

The state-of-art biometric systems allow for multiple users – up to 5,000. Apparently an e-commerce application may have many more. Whether such a system can scale up is not yet quite clear. There is a physical limitation here. The more increased the inter-speaker sensitivity (in other words, the more users the system can distinguish), the more likely becomes for that system to be affected by intra-speaker variations (the more the probability of recognition errors for the same person). Perhaps, biometric systems should rely on passwords *and* voiceprints, which necessitates remembering passwords but still eliminates the use of keyboards. It is indicative that Microsoft has decided to support biometric technologies in the future editions of its operating systems [H40].

5. Conclusions

In this paper issues regarding the speech-enabled e-commerce for the elderly and the disabled user groups have been critically reviewed. Following the last part of the paper, one could say that speech-enabled e-commerce is important to ordinary users. Speech technologies may contribute to enhance and improve existing interfaces on different platforms including common and telephone terminals and information kiosks. Speech technologies are even more important to many elderly and disabled since they are unwilling or unable to use the typical visuo-motor interaction paradigm. E-commerce is even more important to many users in these groups, since it is a matter of access and hence necessity rather than only of convenience. In a nutshell, speech-enabled e-commerce has plenty to offer to the elderly and the disabled, perhaps even more than it has to ordinary users.

Perhaps the most serious problem is to become aware that these users represent a substantial and increasing proportion of the population; to increase the awareness of their needs. This paper has demonstrated how they could be benefited by speech technologies and why these technologies are even more important to them than to ordinary users. It is quite promising that protocols and standards are under rigorous development and that in the immediate future speech-enabled e-commerce will become a reality. Speech technologies are becoming more and more mature and adequate for most e-commerce constituent applications, as the present paper has demonstrated. There is substantial effort expended and a convergence in the different

research frontiers related to speech-enabled e-commerce and mobile commerce, which are no more a remote possibility but very close to becoming reality.

There are still some issues open though, regarding robust speech recognition and high quality speech production, as well as the processing power of mobile devices and the finalization of protocols like WAP to support voice browsers and speech technologies in general. Furthermore, many human-computer interaction issues have to be resolved, to achieve powerful and effective user interfaces; one cannot merely transform a GUI into a dialog system in a superficial way. E-commerce issues have also to be refined – for instance refer to projects like MKBEEM [H45], SPOTLIGHT [H46] and NESPOLE! [H47].

References

Boothroyd, D. (1999). Phoning up the Web. Le Journal, the journal of record for human language technology, [H41]

Boyce, S. (2000) Natural Spoken Dialogue Systems for Telephony Applications, Communications of the ACM, Vol. 43, No 9, Sept. 2000, pp.29-34.

Brewster, S. A. (1995). Providing a Structured Method for Integrating Non-Speech Audio into Human-Computer Interfaces. Doctoral Thesis, Department of Computer Science, York, University of York, UK.

Corredor-Ardoy, C., Gauvain J. L., Adda-Decker, M., and Lamel, L. (1997). Language identification with language-independent acoustic models. In Proceedings of Eurospeech'97, Rhodes, Greece. Vol. 1, pp. 55-58.

Cunningham, P. and Fröschl, F. (1999). Electronic Business Revolution. Opportunities and challenges in the 21st Century. Springer, 1999.

Deshmukh, N., Ganapathiraju, A. and Picone, J. (1999). Hierarchical Search for Large-Vocabulary Conversational Speech Recognition. IEEE Signal Processing Magazine, September 1999, pp. 84-107.

Hawley, M. S. (1999). Automatic Speech Recognition and People with Severe Physical Disabilities. In Proceedings of Talking to Computers II, Sheffield, July 9, 1999.

Kouroupetroglou, G. and G. Nemeth (1995): "Speech Technology for Disabled and Elderly People", chapter in the book "Telecommunications for All", Ed. Patrick Roe, Published by the European Commission - Directorate General XIII, Catalogue number: CD-90-95-712-EN-C, 1995, pp.186-195.

Lai, J. (2000) Conversational Interfaces, Communications of the ACM, Vol. 43, No 9, Sept. 2000, pp.24-27.

Lucas, B. (2000) Voice XML for Web-based distributed Conversational Applications, Communications of the ACM, Vol. 43, No 9, Sept. 2000, pp.53-57.

Lucente, M. (2000) Conversational Interfaces for e-commerce applications, Communications of the ACM, Vol. 43, No 9, Sept. 2000, pp.59-61.

Markowitz, J. (1996). Using Speech Recognition. Prentice Hall PTR, New Jersey, 1996.

Markowitz, J. (2000) Voice Biometrics, Communications of the ACM, Vol. 43, No 9, Sept. 2000, pp.66-73.

Mitsopoulos, E. N. (2000). A Principled Approach to the Design of Auditory Interaction in the Non-Visual User Interface. Doctoral Thesis, Department of Computer Science, York, University of York, UK.

Noonan, T. (1999). Accessible e-commerce in Australia: A discussion paper about the effects of electronic commerce developments on people with disabilities. SoftSpeak Computer Services & Blind Citizens Australia. [H42]

Pitt, I. (1998). The Principled Design of Speech-Based Interfaces. Doctoral Thesis, Department of Computer Science, York, University of York, UK.

Roe, P. R. W., Sandhu, J. S., Delaney, L, Gill, J. M. and Mercinelli, M. (1995). Consumer overview. Chapter 2.1 in "Telecommunications for All", ed. Roe, P. R. W., Published by the European Commission – Directorate General XIII (COST 219), catalogue number CD-90-95-712-EN-C, 1995.

Stephanidis C. (ed.), Salvendy G., Akoumianakis D., Arnold A., Bevan N., Dardailler D, Emiliani, P.L., Iakovidis I., Jenkins P., Karshmer A., Korn P., Marcus A., Murphy H., Oppermann C., Stary C., Tamura H., Tscheligi M., Ueda H., Weber G., Ziegler J. (1999). Toward an Information Society for All: HCI challenges and R&D recommendations. International Journal of Human-Computer Interaction. Vol 11, No 1, pp. 1-28, [H43].

Stevens, R. D. (1996). Principles for the Design of Auditory Interfaces to Present Complex Information to Blind People. Doctoral Thesis, Department of Computer Science, York, University of York, UK.

Ström, N. (1998). Position Paper for the W3C workshop: "Voice Browsers", Spoken Language Systems Group, MIT Lab for Computer Science, [H44]

Tufte, E. R. (1997). Visual explanations: images and quantities, evidence and narrative. Chesire, CT, Graphics P., 1997.

Zissman, M. A. (1997). Predicting, diagnosing and improving automatic language identification performance. In Proceedings of Eurospeech'97, Rhodes, Greece. Vol. 1, pp. 51-54.

Zue, V., Seneff, S., Glass, J. R., Polifroni, J., Pao, C., HAzen, T. J. and Hetherington, L. (2000). JUPITER: A Telephone-Based Conversational Interface for Weather Information. IEEE Transactions on Speech and Audio Processing, Vol. 8, No. 1, January 2000, pp. 85-96.

References to Hyperlinks

H1 http://www.dragonsys.com/products/naturallyspeaking/standard/index.html
H2 http://www.zdent.com/pcweek/reviews/0504/04lern.html
H3 http://www-4.ibm.com/software/speech/millenium/standard/index.html
H4 http://www.speech.philips.com/ud/get/Pages/h072lus_HEat.htm
H5 http://asru99.research.att.com
H6 http://www.speech.philips.com
H7 http://www.nuance.com
H8 http://www.nuance.com/index.htma?SCREEN=verifier
H9 http://www.zdnet.com/products/stories/reviews/0,4161,386987,00.html
H10 http://www.veritelcorp.com
H11 http://www.intelitrak.com
H12 http://www.sanders.com/spard/voice/v_li.html
H13 http://www.speechworks.com/demos/index.cfm
H14 http://www.att.com/aspg
H15 http://www.bell-labs.com/project/tts/voices.html
H16 http://www.eloq.com
H17 http://www.apple.com/macos/speech
H18 http://www.hj.com/JAWS/JAWS.html
H19 http://www.dolphinuk.co.uk/products/HAl1.html
H20 http://www.zdnet.co.uk/pcmag/trends/1999/06/realspeak.html
H21 http://www.lhs.com
H22 http://www.techweb.com/wire/story/reuters/REU19990506S0003
H23 http://www-4.ibm.com/software/speech/telephony/index.html
H24 http://www-4.ibm.com/software/speech/crm/universal_access.html
H25 http://www.crossmedia.net/pages/demo.html
H26 http://www.sls.lcs.mit.edu/sls/wHAtwedo/index.html
H27 http://www.microsoft.com/IIT/OnlineDocs/intro2SAPI.html
H28 http://java.sun.com/products/java-media/speech/forDevelopers/JSML/
H29 http://java.sun.com/products/java-media/speech/forDevelopers/JSML/JSML.html
H30 http://www.voxml.com
H31 http://www.voxml.com/downloads/VoxMLwp.pdf
H32 http://www.voicexml.org/pr20000307-1.html

H33 http://www.w3.org/Voice
H34 http://www.w3.org/Voice/TalkML
H35 http://www.v-commerce.org/
H36 http://www.mobileipworld.com/wp/whitepaper.html
H37 http://www.wapforum.org
H38 http://www.w3.org/2000/09/Papers/Agenda.html
H39 http://www.mobilemexe.com
H40 http://www.in.gr/innews/narticle.asp?nid=18238
H41 http://www.lingling.lu/LeJournal/article.asp?articleIndex=2055
H42 http://www.bca.org.au/ecrep.htm
H43 http://www.ics.forth.gr/proj/at-H36i/files/white_paper_1999.pdf
H44 http://www.w3.org/UI/Voice/1998/Workshop/MIT-SLS.html
H45 http://www.linglink.lu/hlt/projects/mkbeem/
H46 http://spotlight.ccir.ed.ac.uk/
H47 http://nespole.itc.it/