CLASS Natural and Multimodal Interactivity Deliverable D1.7

Natural Interactivity Report on Emerging Market Opportunities

February 2003

Authors

Niels Ole Bernsen and Manish Mehta
Natural Interactive Systems Laboratory, University of Southern Denmark

Contents

NATURAL INTERACTIVITY

Report on emerging market opportunities

Niels Ole Bernsen and Manish Mehta

Summary

This CLASS (Collaboration in Language and Speech Science Technology) report discusses the emerging market opportunities for natural interactive communication systems technologies. The vision of natural interactivity is presented and related to the idea of multimodal systems. The components of natural interactivity are presented. As full natural interactive communication systems technologies are future technologies which presuppose that numerous research challenges will be solved successfully, the natural interactivity vision is operationalised into the more limited emerging technology vision of domain-oriented natural interactive systems. A number of time-lined key research challenges to domain-oriented natural interactive systems are presented followed by a market strategic view of current and emergent market opportunities in the field. For a more in-depth analysis of current research needs in natural and multimodal interactivity engineering, the reader is referred to [Bernsen and Dybkjær 2003]. The present report is an expanded and updated version of the draft CLASS report with the same title. The update takes into account deliverables from our CLASS partners as well as results of the CLASS Verona Workshop on Intelligent Interactive Information Representation [Bernsen and Stock 2001] and the CLASS Copenhagen Workshop on Natural, Intelligent and Effective Interaction in Multimodal Dialogue Systems [van Kuppevelt et al. 2002].

1. Introduction

Natural interactivity refers to a vision of the future of interaction between humans and computer systems. The vision is technological in nature and states the goal to eventually produce systems with which users communicate in the same ways in which humans communicate with one another. This goal is evidently of a very long-term nature, so long-term, in fact, that it makes little practical sense to investigate the market opportunities it affords at the present time. However, it is an interesting property of the natural interactivity vision that we are intimately familiar with the ultimate goal it proposes, and that this goal generates a series of well-defined research challenges which must be met in order to reach that goal. By consequence, it is possible to operationalise the goal in terms of a "first relay" goalpost which must be passed on the way towards full natural interactivity. This goalpost is called *domain-oriented natural interactive communication* below. Given that goalpost, it becomes possible to (a) define the minimum number of research challenges which must be overcome to reach it, and (b) inspect the strategic commercial start and end opportunities in the run towards the goalpost. This is what the present report is about.

In the following, Section 2 defines natural interactivity by contrast with GUI human-system interaction and discusses the potential of natural interactivity for radically improving interaction. Section 3 relates natural interactivity to the wider and somewhat less focused idea of multimodal interactivity. Section 4 presents the behavioural components of natural interactivity. Section 5 provides a working definition of the intermediate goalpost of domain-oriented natural interactive communication (DNI) and goes on to present nine outstanding research challenges which have to be met to achieve DNI. Section 6 looks at the state of the art in current DNI “predecessor” technologies. Section 7 presents findings on key market areas.

2. Natural interactivity, vision and promise

Natural interactivity represents a technological vision which is primarily aimed at overcoming or, at least, massively improving our available solutions to, a basic problem in human-computer interaction (HCI). The problem is that, despite decades of interdisciplinary effort to make computer interfaces easy to use and intuitively accessible to humans, computer interfaces remain largely opaque and non-intuitive. This struggle has involved a large number of disciplines, such as psychology, sociology and ethnography as well as systems developers and software engineers. Still, the prevalent graphical user interface (GUI) paradigm, involving mouse, keyboard, and screen continues to generate computer interfaces which conceal system functionality to the extent that, say, an average 90% of the system's functionality remains forever ignored by its users. The reason people cope is that they have to, at work, with the VCR they have bought at home, with their mobile devices, their palm device, etc. Even if these latter devices add human-to-human speech and handwriting input to the input/output modalities available to their users, the device interface itself remains a GUI. The principal way in which people cope is to ask colleagues or friends how to operate those systems and devices, learning the absolutely necessary minimum from other people in order to do at least part of what they are supposed to do or wish to do. Most of those who read the manual find that they are misdirected as often as helped by it.

It seems fair to conclude that GUI interfaces, although existing in, literally, billions of devices across the world, do not even come close to releasing the full potential of information and telecommunication technologies: they are incapable of releasing the full potential of the devices to which they constitute the interface; it is easily possible to conceive of large families of applications for which it does not make much sense to use them; and for many users, such as the +one billion illiterates in the world and many disabled groups, these interfaces constitute a formidable obstacle to benefiting from technological progress. It should be noted that an "interface" to a system is the user-system meeting-point at which the user or, increasingly, the users, exchange information with the system in order to make the system do what the user wants it to do. In other words, GUI interfaces are seriously flawed solutions to the problem they are meant to solve, that is, the problem of facilitating rather than providing obstacles to the communication between the user(s) and the system.

The vision of natural interactivity aims to eventually produce systems with which users communicate in the same ways in which humans communicate with one another. Eventually, such interfaces will transform systems into quasi-human interlocutors which find it as easy, or as difficult, as the case may be, as humans do to interpret people's messages and behaviours and transform these into actions desired by the users.

With natural interactivity, gone will be the steep-but-stifled learning curve which GUI systems users have to go through to operate computer systems at a minimum level of proficiency. Secondly, all users will benefit because natural interactivity will not prevent the illiterate or the disabled from using computer systems. A third promise of natural interactivity is to release the enormous potential of computer systems to help users with all kinds of task and beyond, a potential of which we are currently only seeing small fragments due to the prevalence of GUI systems. The reason why natural interactive systems can help achieve large families of applications which are not possible today is the vastly improved communicative expressiveness of natural interactive interfaces compared to the use of current designer-created, menu-based graphical output interfaces through which users have to tell the system what they want it to do.

In other words, natural interactivity promises not only ease and intuitiveness of use for all people but also new generations of systems and devices which can do what cannot be done by present systems. However, this claim remains a wish for the cavalry to come to our rescue from the GUI onslaught as long as we have not mastered the natural interactivity research challenges (Section 5).

In a sense, the natural interactivity vision is as old as the computer. Following the invention of the general-purpose computer towards the end of the 2nd World War, it was quickly envisioned that computers might in principle equal or even surpass human intelligence. In the emerging Artificial Intelligence (AI) visions at the time, equality became identical to computers which could master human language and hence pass the Turing test [Turing 1950] by deceiving people into believing that they were having conversation with another human rather than with a computer. In the natural interactivity vision from the late 1990s, however, language has been replaced by the full arsenal of modalities for information representation and exchange which humans use during communication, arguably making it even harder for machines to pass this new natural interactivity test than to pass the classical Turing test.

3. Natural interactivity and multimodality

It is useful to briefly compare natural interactivity and multimodality in terms of human-system interface properties. It is always arguable how far back a particular computing paradigm may be traced. An early paper in the emergence of natural interactivity is [Bolt 1980]. A corresponding paper on multimodality which prompted one of the authors to start investigating the field is [Hovy and Arens 1990].

In the last couple of years, the term 'multimodal system' has suddenly become commonplace in the computer and IT world. A *modality* is simply a way or manner or way of presenting or exchanging information at the human-computer interface or among humans. Thus, *multimodal systems* are systems which use several modalities for this purpose. Their component modalities are called *unimodal modalities* for information (re-)presentation and exchange. Examples of unimodal modalities at some relatively low level of abstraction are graphical images, written text, speech, non-speech sounds, data-graphics, haptic or touch (Braille) written text, etc. It is very important to distinguish between *input* (from the user) and *output* (from the system) modalities, both for clarity of exposition in general but also because, generally speaking, a certain output modality tends to first appear in practically useful system output and only later in practically useful system input. *Modality Theory* is the theory of the properties of modalities [Bernsen 1994, 2002]. Thus, the multimodality idea resembles the natural interactivity vision in the call to go beyond GUI interfaces in order to improve human-system interaction and enable so far infeasible families of applications. More specifically, there is general agreement that multimodal interfaces could offer the following advantages [Benoit et al. 2000]:

1. modality synergy;
2. different modalities, different benefits;
3. increased expressiveness;
4. new applications;
5. freedom of choice;
6. naturalness, little or no learning overhead;
7. adaptation to different environments;
8. users with special needs.In this list, (1) *modality synergy* refers to the fact that several modalities may cooperate to get a message across in a more robust manner, such as the simultaneous use of spoken output and (synchronous) lip movements (see also below). (2) *different modalities, different benefits* is self-explanatory. (3) *increased expressiveness* and (4) *new applications* were mentioned above. (5) *freedom of choice* refers to the possibility for users to choose which modalities they want to use for exchanging information with the system. (6) *naturalness, little or no learning overhead* was mentioned above. (7) *adaptation to different environments* refers to the fact that different modality combinations may be appropriate for different environments even for the same application. For instance, graphics output text may be preferable to spoken output in noisy environments or in environments in which speech might disturb other people. (8) *users with special needs* was mentioned above.

As will be expanded on below, natural interactivity is multimodal most of the time. So, the two main differences between the multimodality idea and the natural interactivity vision would seem to be (i) that multimodality could include modalities for information representation and exchange which are not being used among humans, for instance due to the limitations of their perceptual systems, such as ultrasound or infrared; and (ii) that natural interactivity constitutes a structured vision with a well-defined goal whereas multimodality is simply the idea of trying out what is in fact a very large number of possible modality combinations in order to identify those which might serve to improve human-system interaction or enable new applications. In fact, the number of possible input/output modality combinations run into hundreds of thousands most of which are hardly useful at all, and natural interactivity scenario analysis has shown that most modalities are actually being used in human-human natural interaction anyway [Bernsen 2002].

4. Components of natural interactive communication

The components of natural interactive acts of communication are:

- speech;
- facial expression including lip movement and gaze;
- hand/arm gesture including pointing;
- body gesture including head movement and body posture;
- by extension, hand/arm gesture opens up for touching, holding, showing, manipulating and otherwise dealing with objects as part of natural interactive acts of communication. Some of these objects may themselves represent information, such as images or text, whereas others are tools (devices) for creating information representations, such as the keyboard, the mouse, or the pen, and others are neither of those things.

All of the above components are, or can be, integral parts of a particular natural interactive act of communication. When used, the components are often mutually redundant and complementary in the information they convey. *Redundancy,* such as between speech and lip movement, makes for robust communication, reducing the risk of miscommunication. *Complementarity,* such as between text and images, makes for increased expressiveness (or, metaphorically expressed, "bandwidth), enabling full natural interactive communication to convey virtually any conceivable message. Also, in human-human communication, the components of natural interactivity act as both input and output modalities most of the time.

As simple as the above list of components of natural interactivity may appear, harnessing natural interactive communication for the purpose of human-computer interaction constitutes a tall research agenda for many decades to come.

5. Natural interactivity state of the art

This section presents a brief and condensed state of the art in natural interactivity, focused on a number of key research challenges. An in-depth analysis of current research needs in natural and multimodal interactivity engineering is presented in [Bernsen and Dybkjær 2003].

5.1 Definitions

Today, natural interactivity is a vision rather than reality. It will take decades, at least, to achieve full natural interactivity, if, indeed, this will be possible to do at all. For the purpose of the present report which is to look into comparatively short-term market opportunities, it may be useful to distinguish between the following two goals:

- domain-oriented natural interactivity;
- full natural interactivity.

In the following, we focus on domain-oriented natural interactivity, or DNI.

An operational definition of DNI could be the following. DNI will have been achieved when the system understands and generates conversational speech coordinated in real time with practically adequate facial expression, gesture and bodily posture in all domains which require up to, say, 2000 basic word forms. *Practically adequate facial expression, gesture and bodily posture* may not be perfect replicas of human communicative behaviour, but they work, that is, they are being understood by humans as intended in the system's output, and they are being understood by the system as intended by humans when providing input to the system. *Basic word forms* are morphological stem word forms of (i) the most widely used words in a particular language as well as (ii) the words which are characteristic of the domain. Of course, if the domain includes thousands of, e.g., proper names all of which have the same function(s), such as the address items to be input to a car navigation system, the DNI system, assuming powerful speech recognition, could master a vocabulary of 100.000 words or more, since +98.000 of those words could be treated by the system in essentially the same way.

Domain-oriented systems, whether natural interactive or otherwise, constitute a beyond-the-state-of-the-art generation of systems whose capabilities generally go beyond those of task-oriented systems. The term 'task-oriented system' derives from the spoken language dialogue systems (SLDSs) domain [Bernsen et al. 1998] but essentially denotes all existing computer systems. A *task-oriented system* allows the user to accomplish a particular task, or several tasks, such as getting train time-table information over the telephone or composing and formatting a document. Task-oriented systems have proved of tremendous value to the field of spoken dialogue systems (SLDSs), providing the field with its first general *application paradigm,* i.e. a model for a large family of applications of substantial market impact. Task oriented SLDSs have been in the market for more than ten years and continue to develop for new and more complex tasks and in new languages. A *domain-oriented system,* however, is a system which is no longer defined through the task(s) it can help the user accomplish. Rather, a domain-oriented system is defined solely through the domain it knows about. Thus, any topic within the domain can be handled by the system to an (approximation to) arbitrary depth, and the actual topic(s) to be addressed in the conversation depends entirely on the user. From the point of view of domain-oriented systems, task-oriented systems are highly restricted or specialised forms of domain-oriented systems. Moreover, domain-oriented systems are *conversational systems:* they are userindependent and understand free-initiative and spontaneous or free-form natural interactive communicative input from users in their domain(s).

Compared to domain-oriented natural interactive systems, *full natural interactive systems* are the same except for the crucial difference that full natural interactive systems, like humans, are able to carry out conversation in any domain based on a more or less deep knowledge-base. In addition, full natural interactive systems would require full and adequate emulation of both verbal and non-verbal human natural interactive behaviour. Thus defined, full natural interactive systems even exceed the demands of the classical Turing test, indicating that full natural interactivity is a very long-term goal. More practically speaking, once DNI has been achieved, we will already be in the process of going beyond it towards multi-domain systems, systems with improved speech-facial-gesture coordination, systems with even more human-like communicative behaviour, etc.

5.2 State of the art

The following is a brief look into some of the key research challenges posed by the natural interactivity vision. A talk in March 2001 on large-scale projects for FP5 [Bernsen 2001] identified a minimum of nine research challenges which must be met in order to achieve DNI systems. The scenario used for illustration in the talk was that of having domain-oriented natural interactive conversation on his life and works with the world-famous author Hans Christian Andersen. In the scenario, Andersen is able to conduct domain-oriented conversation with several people at the same time and he freely manipulates objects during the conversation, said objects contributing to what he is talking about in the same ways in which humans use, or refer to, situated objects during conversation. The objects are of the kinds mentioned in Section 4 above, i.e. information representing objects, such as texts and images, and other objects, such as people, animals, houses, trees or landscapes. Figure 1 shows an early in-house version of a young Hans Christian Andersen able to carry out simple dialogues about his life and works. Meanwhile, a limited version of the Hans Christian Andersen project has been launched in 2002. The project is called NICE (Natural Interactive Communication for Edutainment, and focuses on challenges 1, 3, and 5 below, cf. [http://www.niceproject.com/]. Figure 2 shows the mature NICE Hans Christian Andersen who will be having domain-oriented natural interactive conversation with children and adolescents.

The research challenges identified in the above-mentioned talk are:

1. domain-oriented spoken conversation (+SLDSs);
2. multi-party (and multi-lingual) speech recognition and understanding (+CSp);
3. integrated graphical animation, emotion (+CGr), and
4. prosody (+CSp);coordinated presentation of text, images, non-speech sound, object manipulation etc. (+IIIP);interlocutor tracking and identification (+CVi);audio-visual speech recognition (+CVi, CSp);emotion and interest interpretation through face and gesture (+CVi);on a portable platform for web, open mike, and mobile devices (+SSWE).In the above list,

abbreviations in brackets are as follows:

- SLDSs = spoken language dialogue systems.
- CSp = computer speech.
- CGr = computer graphics.
- IIIP = intelligent interactive information presentation.

- CVi = computer vision.
- SSWE = serious software engineering (for integrating complex systems which can handle the solutions to the above research challenges in real time).

What the above list of key research challenges to be overcome in order to achieve DNI shows, is (i) that progress towards DNI requires bridging between a goodly number of islands in the present ITC/software engineering archipelago whose splendid isolation has been upheld throughout the 1990s. Today, researchers in any of the fields listed above know far too little about research in any of the other fields. Moreover, a new culture of cross-disciplinary collaboration among those fields of computer science and engineering is still to develop at this point; and (ii) that progress towards DNI, and towards natural interactivity more generally, has the excellent property of being measurable against the extent to which a series of well-defined research challenges has been met. In other words, the DNI vision, as well as the vision of full natural interactivity, determines a semi-ordered series of research challenges to be addressed.

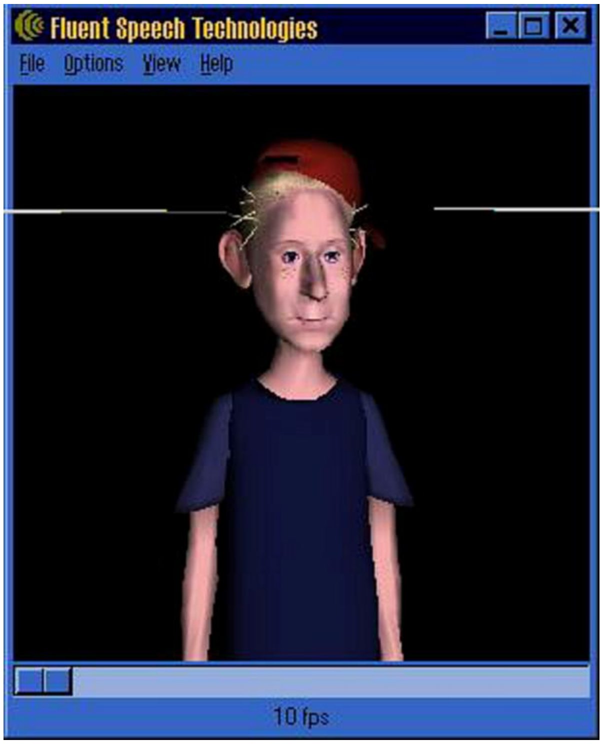

Figure 1. In-house rapid prototype of a young Hans Christian Andersen.

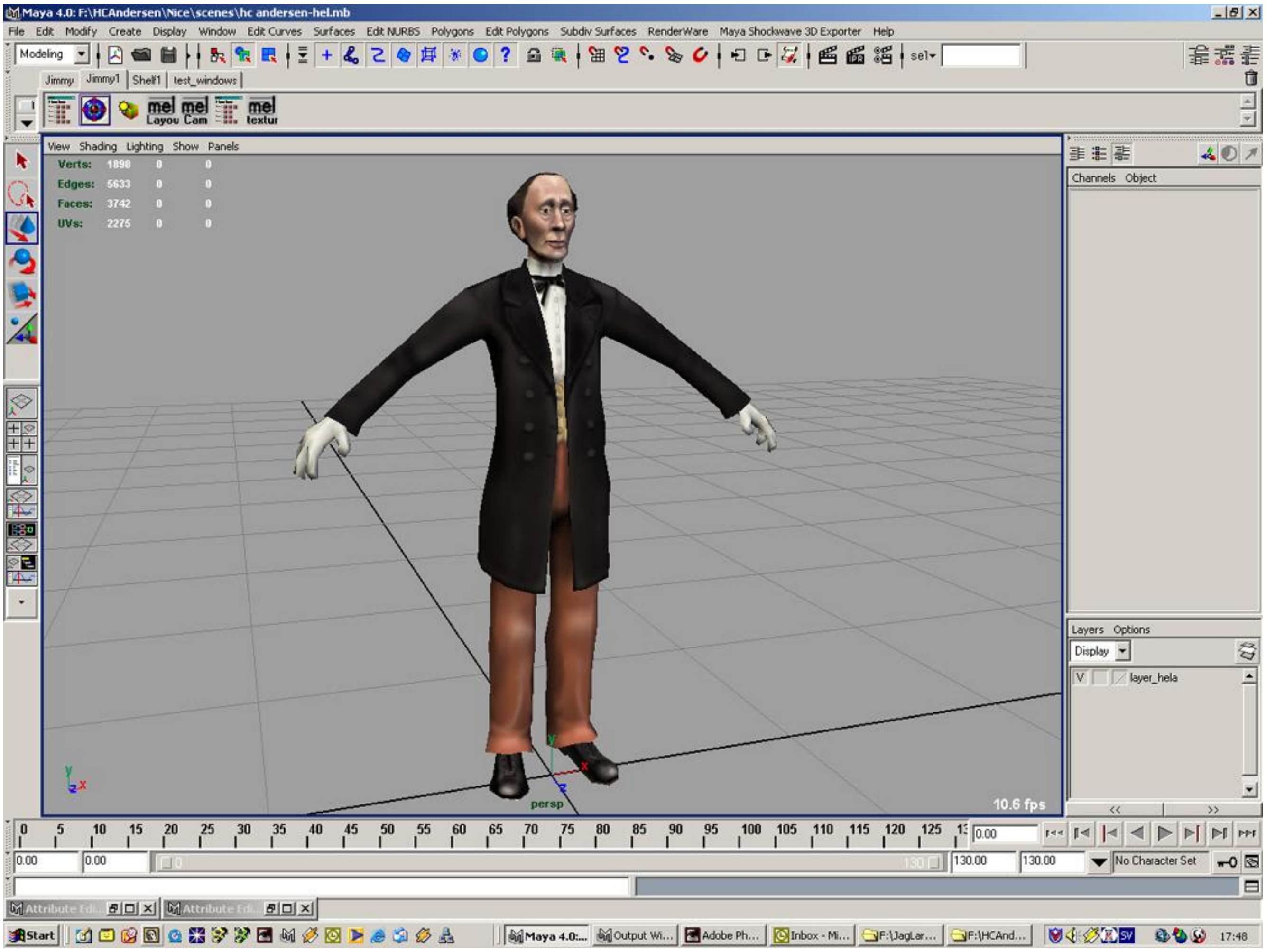

Figure 2. The NICE Hans Christian Andersen in initial position for animation.

Let us briefly look at the state of progress towards meeting the nine research challenges above. The comments below are partly based on an ELSNET (European Language and Speech Network) brainstorming roadmap paper on challenges for the next ten years of research in speech technologies [Bernsen et al. 2001]. Like in the ELSNET roadmap, estimated times refer to when, assuming adequate effort, solutions ready for industrial development have been achieved in research.

1. *Domain-oriented spoken conversation.* A breakthrough is needed for building such systems. The most advanced task-oriented SLDSs which are being built by researchers today, such as the in-car system which is being developed in the European VICO (Virtual Co-driver, www.vico-project.org) project, aim to handle multiple tasks speaker-independently in noisy conditions with free-initiative user speech and 100.000 words vocabularies. Compared to these systems, the key problem to be solved in order to achieve domain-oriented spoken conversation is to generalise systems capable of handling VICO-style communicative complexity so that the system is no longer constrained by any task structure. This requires new methods of knowledge representation and dialogue modelling. The estimated time for demonstrating a working, DNI-compatible demonstrator solution to the problem is estimated at 3 years.

2. *Multi-party (and multi-lingual) speech recognition and understanding.* Some points on the technical challenges of multi-linguality are made in [Bernsen and Dybkjær 2003]. Multi-party speech recognition and understanding is a hard problem. Reflecting the importance of solving this problem, US NIST (the National Institute of Standards in Technology) and US DARPA (Defence Advanced Research Projects Agency) recently have begun to address it through the scenario of meeting transcription. NIST has built a meeting room replete with cameras and microphones in order to create data corpora for developing and testing multi-party speech recognition systems. The estimated time for demonstrating a working solution to the problem is 2-3 years.

An interesting, related point concerns the "for all users" or *universality* part of the natural interactivity vision. It is a perhaps slightly surprising fact that current speech recognition technologies, in addition to being incapable of handling multi-party speech recognition at a level of performance suitable for product development, generally perform less well than desired when it comes to recognising the speech input produced by children and young people. Specialised recogniser training as well as research on the spoken linguistic behaviour of children and young people will be needed to overcome this problem. The time-scale for solving the problem is estimated to 2 years.

3. *Integrated graphical animation and emotion.* 2D and 3D computer graphics are already at a highly advanced stage of development, for instance in the computer games industry. Techniques from the computer games industry have been applied to the generation of life-like animated characters for a long time already. The field of animated (graphical) interface agents is in rapid development to the extent that an MPEG7 standard exists for distinguishing between "the seven major" human emotions in graphical output agent design. The estimated time for demonstrating a working solution to the problem for DNI is 2 years. It should be added, however, that although the solution might comply with the operational definition of DNI above (Section 5.1), it might also not do so.

Moreover, the issue of producing *practically adequate facial expression, gesture and bodily posture* is far from being merely an issue of solving the problem of facial emotion generation. The human face is extremely expressive, and part of its expressiveness is related to speech behaviour (cf. Point 4 below). We are far from being able to adequately model the expressiveness of the human face and of human gesture and their relationship to speech behaviour, progress being hampered by such basic needs as appropriate natural interactive communication corpora for research as well as appropriate coding schemes and tools for annotating those corpora. The field of natural interactivity corpora, coding schemes, and coding tools is about to emerge as a very important one, as evidenced by the natural interactivity corpora, coding schemes, and coding tools survey work carried out in, e.g., the EU Natural Interactivity and Multimodality (NIMM) Working Group of the EU-US project ISLE (International Standards for Language Engineering, isle.nis.sdu.dk), the EU HLT project for developing a world-first coding tool for natural interactivity corpora, NITE (Natural Interactivity Tools Engineering, nite.nis.sdu.dk), or the natural interactivity coding and development work carried out in the German SmartKom project (smartkom.dfki.de), cf. also [Bernsen and Dybkjær 2003].

4. *Integration of (3) above with prosody.* This problem remains a thorny one because of the facts that speech prosody input/output (i) is extremely multifaceted in terms of the aspects of the speech signal used, (ii) depends on a range of non-linguistic factors, such as personality, the social environment of the discourse, and culture, (iii) depends on a range of properties of the linguistic context which remain incompletely understood, and (iv) probably to some extent has to be understood in the wider context of natural interactive communicative behaviour, including facial expression, gesture and bodily posture.

Demonstrating a working solution to the problem could take 5-10 years, especially since results will have to be integrated with the results of the work described under Point 3 above.

5. *Coordinated presentation of text, images, non-speech sound, object manipulation etc.* Part of this challenge was addressed already about a decade ago in research demonstrator systems, such as those of [Feiner and McKeown 1993], and the coffee machine instruction system of [André and Rist 1993], as well as through the development of a proposed standard reference architecture [Rist el al. 1997]. So far, however, this work, largely has remained a series of research monuments rather that accelerators of development in the field of Intelligent Interactive Information Presentation (IIIP). Today, we need generalisable and practical solutions to the problem which is moving towards centre-stage due to the proliferation of graphical interface agent applications (cf. Point 3 above). What is needed is architectures and generalisable demonstrators which show how animated graphical interface agents can be combined with the kinds of use of objects during communication which humans employ. The problem is clearly a solvable one, and it may be expected that the need to give animated interface agents something (more) to do, such as objects to manipulate, will constitute a major driving force towards its solution. The estimated time frame for reaching a working solution to the problem is 3-10 years, depending on the complexity of the interface agents' behaviour.

6. *Interlocutor tracking and identification.* Computer vision researchers have some of the hardest challenges to overcome in the present list. Among these challenges, however, camera-based interlocutor tracking and identification are already within reach, as witnessed by commercial surveillance systems in the UK and the US. Moreover, the current push towards smart camera surveillance systems development following 11 September 2001 seems likely to accelerate progress in the field, the results of which can be put to other, less controversial, uses than surveillance. The estimated time for demonstrating a working solution to the problem is 2-3 years.

7. *Audio-visual speech recognition.* This is another area of rapid progress as it has been found that combined acoustic/visual speech recognition improves speech recognition especially in noisy conditions. The estimated time for reaching a working solution to the problem is 2 years. On the output side, audio-visual speech recognition is mirrored by realistic, real-time synchronised generation of combined speech synthesis and lip movement, an area which forms part of Points 3 and 4 above.

8. *Emotion and interest interpretation through face and gesture.* Among the problems in the present list, this cluster of interrelated issues is probably the most difficult one. Solutions have to assume significant progress, or even breakthroughs, in computer vision, and those solutions depend as much on advance in natural interactivity corpora, coding schemes, and coding tools as do the corresponding output generation issues (Points 3 through 5 above). The computer vision community is presently gearing up to address these challenges and to build the bridges to other computer science disciplines which are necessary for accelerating progress. Again, current needs for improved camera-based surveillance systems is likely to give the field an initial push. Moreover, indications are that the computer vision community is about to go beyond person tracking and identification to the visual interpretation of behaviour, emotion and character. This was clearly borne out when one of the authors gave an invited talk at the recently established FGNet (Face and Gesture Network) workshop in Grenoble in November 2001. FGNet is a network of European computer vision researchers and it was clear from the brainstorming exercises done at the workshop that their research agenda is about to comprise elements of natural interactivity. The estimated time frame for demonstrating a working solution to the problem is 5-15 years.

9. *Natural interactivity on a portable platform for web, open mike, and mobile devices.* The base technology challenges which must be overcome to solve this cluster of problems are likely to have been met within 3-4 years.

In conclusion, it would seem clear that we can expect the coming ten years or so to demonstrate dramatic progress in natural interactivity technologies. Given the estimated time scales above, and given the fact that component development must be complemented by integrated real-time demonstrator systems development, it seems fair to estimate that domain-oriented natural interactivity can be achieved in less than 10 years. From an industrial/commercial point of view, however, it is of equal importance to note that the DNI future starts today. In view of the time scales just presented, it would seem clear that limited-capability DNI systems, i.e. systems which incorporate early progress in one or several of the nine research areas discussed, could be specified for research prototype development right here and now in order to position first products in the market at the earliest possible.

6. DNI market opportunities

6.1 DNI predecessor technologies and competitors

As is almost always the case when describing a vision about the future of computing, the future described has already begun in some sense and to some degree. This is also the case for DNI applications. Even though none of the research challenges described in Section 5 have been solved yet, predecessor technologies already exist which provide a sense of what is to come. Analysis of those predecessor technologies would seem crucial to the identification of DNI market opportunities from at least two different points of view:

- knowing what today's state-of-the-art DNI-related technologies can do already is a useful way of identifying the next commercial steps towards DNI applications; and
- knowing the competitive landscape of DNI predecessor technology suppliers is essential to the development of a strategy for positioning one's company in the market.

Clearly, the owners of DNI predecessor technologies have a smaller or larger advantage over newcomers, whether these are new start-up companies or established companies wanting to branch into DNI technologies and applications. Companies which already produce those technologies could be well positioned to take on a smaller or larger part of the DNI challenge. Similarly, new companies who would like to take on the DNI challenge may be well advised to carefully consider what is already there, who produces it, and how to specify the best possible company launch platform in the landscape described. Finally, whether DNI newcomer or predecessor technology producer, companies need a clear view of the state of the art in order to identify their next technology development steps.

The following provides an overview of DNI predecessor technologies and the next steps they could enable.

1. *Domain-oriented spoken conversation.* In general, spoken dialogue is fundamental to natural interactive communication and is expected to constitute a key capability of most DNI-related technologies. Spoken language dialogue systems companies are well positioned to address DNI, especially those forefront organisations who already produce medium-vocabulary, speaker-independent, mixed-initiative task-oriented applications. An obvious next step is to combine these technologies with state-of-the-art animated interface agent technologies. The challenge is to identify first "killer applications" for this combined technology, of which none seem to have been forthcoming yet. For the most part, existing research demonstrators combining task-oriented spoken dialogue and animated interface character technology appear as technology in search of meaningful applications rather than demonstrations of a new powerful application paradigm. Tutoring applications have become popular in US research [Bernsen and Dybkjær 2003] but, so far, at least, seem to a large extent to rely on "big money" sponsoring from the military. It is far from clear when a breakthrough into the educational mass market could happen. We would rather like to suggest to consider the surprisingly underexplored mass market for edutainment and entertainment applications of spoken dialogue and animated interface agent technologies, and primarily the market for children and young people, as spearheaded by, for instance, the NICE project [http://www.niceproject.com/]. The hybrid expression 'edutainment' refers to applications which educate and entertain at the same time. The world of cultural communication, including museums and many other kinds of exhibits is a paradigm target for edutainment applications, cf. [Stock and Zancanaro 2002].

2. *Multi-party (and multi-lingual) speech recognition and understanding.* By contrast with spoken language dialogue systems developing organisations of which there are probably hundreds world-wide, all of which have to deal with increasingly complex speech understanding, *speech recognition* technologies are becoming a specialised business area with relatively few strong players world-wide, such as Nuance, SpeechWorks, Scansoft/Philips, and others. Some of these companies might emerge as winners in the conquest for multi-party speech recognition products and speech recognition products for children and adults. The winning companies are well positioned to deliver components to DNI applications as these emerge.

In multi-party speech recognition, various research projects have been going on in different research environments. For instance, the Meeting Recorder project [http://www.icsi.berkeley.edu/~dpwe/research/mtgrcdr/] aims to develop technologies that would enable a palm-sized device to make a useful record of a meeting, through speech recognition, automatic segmentation and information retrieval technologies. At ICSI, they have equipped a meeting room with a multichannel, studio-quality recording system and have begun to collect pilot recordings of meetings, primarily between speech group members, see the information on Meeting Recorder data collection. The purpose of the Robita project [tk.elec.waseda.ac.jp/robita/] is to create a "conversation" system, which can participate and take part in the group conversation. Group conversation is a form of multi-party conversation in which participants can see each face and hear each voice mutually. Group conversation raises many new problems which are not addressed in conventional one-to-one conversation: recognition of message exchange (recognise who is speaking and to whom s/he is speaking), expression of messages to users, and strategies for taking part in the conversation. FOG, developed by CoGenTex, produces two distinct varieties of weather bulletins, public and marine forecasts, in both English and French. The system's input is a forecast map data produced by an atmospheric-modelling program that contains values for predicted air pressure, humidity, temperature, and so forth.

3. *Integrated graphical animation and emotion.* Computer graphics is a strong and well-established business area with thousands of producers and specialised products. Among the producers, computer games companies are perhaps among the better positioned organisations for exploiting emerging DNI market opportunities. It is an interesting fact that, despite a booming computer games industry, no computer game company appears to have begun to combine graphics computer games with interactive speech technologies including simple spoken language dialogue systems. Once speech recognition products for children and adults have emerged, the introduction of pre-DNI (task-oriented) spoken dialogue in computer games would appear an attractive market opportunity. Two companies which are among the early adopters of this view are Philips/Scansoft and the Swedish computer games company Liquid Media, both partners in the NICE project. Since such products are still to emerge, it is impossible to claim more than evident opportunity at this stage. In 10 years, X percent of the computer games sold world-wide will include something akin to DNI. If X = 50%, the market will be huge indeed.

4. *Integration of (3) above with prosody.* Similar to the speech recognition industry but less so, a pattern of strong players may be emerging in commercial speech synthesis technologies, such as Elan, Svox, Infovox, Bell-Labs, Loquendo, Scansoft, and AT & T. However, text-to-speech technologies are only about to spread more widely in the mass market for low-end task-oriented spoken dialogue systems because application developers have had to wait for products having a voice quality which is acceptable to the general user population. Moreover, given the many unsolved research problems in the field of prosody, clear opportunities exist for newcomer companies to start from the front by

specialising in advanced prosody synthesis. This opportunity may even exist in the speech recognition sector, i.e. for advanced prosody recognition. One obvious next step for prosodic synthesis technology developers could be to combine advanced speech synthesis with state-of-the-art animated interface agent technologies in computer games and elsewhere. It is already possible to deploy such applications over the Web, as witnessed by, e.g., animated interface newsreading agents.

5. *Coordinated presentation of text, images, non-speech sound, object manipulation etc.* Products already exist which, e.g., combine text-to-speech with text, images, non-speech sound, etc. By contrast with the areas discussed so far (points 1 through 4 above), the intelligent interactive information presentation field would appear to be a highly unstructured one at the present time. In addition, it seems difficult to guess where it will go in the future as well as how fast it will develop. IIIP of the kind just illustrated is already possible over the Web, but it seems unclear what might be the driving forces. Company product advertising could be such a driving force. However, more sophisticated applications would seem to have to await progress in DNI more generally, including progress in automated language (text) generation, a field with important research challenges of its own which has not been emphasised above but rather considered part of the cluster of challenges to be addressed in spoken language dialogue systems technologies.

Hypertext applications illustrate early efforts in IIIP. Project Reporter was developed in 1993, and later updated in 1999, at CoGenTex to automatically generate a periodic textual report directly from a database. The idea behind Project Reporter came about from the observation that large projects that needed to coordinate work on multiple tasks and subprojects, often maintain databases on expenditures of resources, such as labour hours and money, in order to allow project managers to track the progress of a project and compare it with the initial project goals. As it grew in popularity, Project Reporter was adapted to generate hypertext reports from the standardised databases of commercial project management software. PLANDOC, one of the most advanced report generation applications was developed at Bellcore Research in 1994, PLANDOC was designed to generate summaries of telephone network planning activity. The system has been used by Southwest Bell telephone-operating companies to write the first draft of documents that engineers must file to justify their decisions [Alfaro 2001]. Similarly, systems that are able to provide assistance to customers that face, for instance, technical difficulties with an electronic product, can translate into significant savings for businesses, while also providing high-quality and tailored assistance to their customers. A number of interactive courses also exist that use adaptive hypermedia techniques to personalise learning. For example, the course on Hypermedia Structures and Systems taught by Professor Paul De Bra at the Eindhoven University of Technology in Holland employs adaptive hypermedia instead of a traditional textbook [De Bra, 2001].

6. *Interlocutor tracking and identification.* This is an emerging commercial field which is likely to move more rapidly as a consequence of 11 September 2001. In research, and in addition to obvious surveillance and authentication applications, the technology is already being targeted for more friendly applications, for instance in the home. This field may be considered key to the launch of the first DNI systems which use camera-based information about their interlocutors. Companies which already market camera-based surveillance technologies will of course have an advantage in developing these applications, but the commercial immaturity of the field as a whole means that market opportunities can be exploited by newcomers.

7. *Audio-visual speech recognition.* Audio-visual speech recognition could become a new component technology which will partly replace speech recognition technologies. Novel and non-traditional ASR methods are necessary in order to approach human levels of performance and for speech to become a

truly pervasive user interface. Audio-visual speech recognition is one such source of making large improvements in high noise environments. Vehicle manufacturers around the world are developing and experimenting vehicle navigation systems that use speech to replace manual input of control commands. Noise is a very serious problem for such application due to the interior noise, noise from the car engine, music and other traffic. Visual speech information is not affected by acoustic noise. If successful, audio-visual speech recognition could spread from niche applications to become a key component technology in its own right. If we briefly compare with audio-visual speech synthesis, the first emerging applications of audio-visual speech synthesis target special user groups, such as the hard-of-hearing and language learning for autistic children [Bernsen and Dybkjær 2003].

8. Emotion and interest interpretation through face and gesture. Market opportunities in this challenging field would seem to be open to all. For instance, advanced user attitude interpretation would be useful for adapting interactive feedback in an intelligent tutoring system when it is important to know the user's degree of interest in the information that is displayed or when s/he is interacting in a virtual environment. Automatic customer satisfaction evaluation and detection of emergencies are two other areas where opportunities can be explored.

9. *Natural interactivity on a portable platform for web, open mike, and mobile devices.* Platform and infrastructure issues are extremely important factors in shaping future markets for DNI applications. The reason is that DNI applications are likely to spread across virtually all known communication infrastructures. Some will come with off-the-shelf PC programmes, for instance as spoken dialogue helpers whose domain is how to operate the software. Others will be applications which are contacted over the telephone, mobile phones, PDAs, etc. There will be web-accessed DNI applications, interactive TV DNI applications [Lazzari 2002], etc. Thus, in addition to the much more specialised commercial organisations considered above, such as speech recognition technology providers or computer games companies, large software companies, telecommunication service providers, mobile phone companies, TV companies, film companies, etc. are all likely to take a keen interest in DNI technologies either as customers or as producers. In fact, large companies from several of those sectors are already moving into DNI, developing research prototypes in their own research laboratories. In general, however, these companies are not likely to develop DNI technologies on their own. The reason is the complexity of DNI, which would seem to virtually force, or strongly invite, collaboration between two or even more companies in order to combine the strengths of each. Early examples of this, by anticipation growing trend are VoiceXML [http://www.voicexml.org/] and SALT [http://www.saltforum.org].

6.2 Killer applications

DNI means domain-oriented natural interactive communication. So, it is almost by definition that the DNI killer application will be a domain-competent interlocutor. Even at the generic level, however, there could be lots of those, such as

- the personal assistant in domain X (helper, advisor, problem-solver, organiser, information retriever, etc.);
- the friend;
- the historical or literary character;
- the adversary, e.g. in a computer game;
- the fictitious story character (future, past, humanoid, extraterrestrial, etc.);

- the educator;
- etc.

All of these might develop into killer applications but it is too early to tell. What the list does show is that DNI applications are likely to exist in virtually all areas of computing, from traditional information systems through infotainment and edutainment to entertainment. Moreover, killer applications might just as well be found on the way towards DNI. The best guess as to where remains applications in the mass markets for entertainment and edutainment.

6.3 Between today and DNI

Between today's DNI-related state of the art, on the one hand, and full DNI applications on the other, lies a vast space of opportunity for discovering useful applications including killer applications. Section 6.1 above does not even begin to address this space. Parts of this space, however, is already being addressed in research prototype development projects world-wide, such as in the VICO project which investigates the would-be killer application of free spontaneous speech-based in-car navigation and more, in the NICE project, or in SmartKom, as described above.

7. Key Market Areas

Key speakers at Lang Tech (Berlin 2002) supported natural interactive, intelligent interfaces. Bill Dolan, Head of Natural Language Processing (NLP) at Microsoft, in his keynote address stated that intelligent user interface is the real killer application - which will provide natural conversation with our computers, helpdesks, web browsers, game characters, encyclopaedias, etc. Microsoft plans to develop task-oriented NL interface to web in Windows XP, to allow novice web users to type their goals and requests in natural language. Further challenges are integration of speech technologies, understanding of paraphrases, building a database of common-sense knowledge (Microsoft develops MindNet ontology), document summarisation. Progress is incremental but real, every day millions of users work with language technologies. And the inevitable will happen in '5 years time'. Giovanni Battista Varile from the European Commission, who presented the Sixth Framework Programme for funding EC research, further supported the 'natural interface' theme. Through the IST programme, the EC has a vision of building a knowledge society with ambient intelligence with user-friendly interfaces for all, embedded in everyday objects (e.g. furniture, vehicles, smart objects). Technologies are to be built for the background since the priority is to bring people to the foreground. This is evidenced in a research budget of over 3,600 million euro for Knowledge and Interface Technologies within the IST Programme. In IST Interface Technologies, the key objectives for 2003 are Multimodal interfaces and Cognitive systems. The expectation is that EC will fund under FP6 fewer projects, which are strongly multidisciplinary with many component technologies. Language and speech technologies are considered integral parts of ambient intelligence.

Thus, key players in the market are having natural interactivity as part of their vision for the future. Natural interactivity is going to play a large role in the future of many areas. Let us now briefly look at the key market areas, which will drive businesses in the future. They have been selected on account of the market trends and the strategy adopted by big companies.

7.1 Tourism

According to the Travel Industry Association of America, over 13 Bn. US$ were spent online in 2000 for airline tickets, hotel reservations and car rentals in the US (nearly one third of all online B2C spending in the US in 2000). Online bookings in the US increased by 58% in 2000. For Europe, a massive jump in the number of Europeans logging on to travel websites is reported. The European Internet Travel Monitor reveals that in 2000, 13 Mio. travels in Europe were either initiated or booked via the Internet, which corresponds to 3,5 % of all travels. Specific sites such as travelocity, expedia or the Austrian destination system TIScover are reporting profit already.

In the travel industry we are witnessing total acceptance to the extent that the structure of the industry and the way business is conducted, is changing. Tourists are becoming increasingly demanding in their travel requirements. The net is used not only for information gathering, there is an acceptance of ordering services over the Internet. And it is not the case of only trying one or two services; it is all travel and leisure services. There is a new type of user emerging. The Internet users seem to accept to become their own travel agent, organise their trips themselves and build their own travels and leisure trips. A traveller trying to find a hotel in a city he or she is planning to visit is not generally interested in receiving a full list of all possible hotels, but would prefer to receive a shorter list of hotels that meet his or her required specifications: price range, location, amenities as anticipated in the VICO project.

The tourism domain is an excellent example for the trend towards increasingly personalised services and complex market mechanism providing these services. It reflects that users become part of the product creation process, the trend from customer-focused to customer-driven. Customers start to ask for their personal prices. As an example, take Priceline which uses a reversed auction mechanism where users can define the price they are willing to pay for a product they can partly define, and applying yield management for capacity management. These industry features and user behaviour explain why many companies traditionally outside the tourism field are now entering (or have already entered) this sector. Or take, for example, Microsoft (with Expedia), Bertelsmann, or CNN, in cooperation with T-Online and Preussag, the greatest European travel conglomerates [Werthner 2001].

It is worth mentioning an initiative by the California Division of Tourism, which decided to launch an interactive television system in order to facilitate access to tourism services. "The "California Channel" should be in full operation by 2005 or 2006. By that time, there will be 175 million homes with an interactive capability, according to Caroline Beteta, deputy secretary of the California Division of Tourism and executive director of the California Travel and Tourism Commission. A similar initiative has been announced by Worldspan who is about to launch an interactive television travel-shopping channel that will enable TV viewers to shop for complete vacations packages and other travel directly through their television. Worldspan is providing TV-based interactive travel information to more than 550,000 Charter Communications cable customers across the United States. Worldspan said that this is the first and only global distribution system to offer a travel solution designed specifically for the interactive television space [Lazzari 2002].

7.2 Edutainment

The ideas that learning can be fun, and that fun can promote learning, are transforming attractions that once saw themselves as primarily either education- or entertainment-oriented. This has even spawned a new word: edutainment. The American Heritage Dictionary defines edutainment as "the act of learning through a medium that both educates and entertains." It consists of two equally important parts: the format (entertainment) and the message/content (education) [White 2003]. On the one hand, we are talking about visually spectacular 3D adventure games with interesting characters, plots based on mystery and intrigue, and puzzles that are seamlessly woven into the stories. On the other hand, we are being presented with the opportunity to explore important historical locations, study collections of artworks, plans, maps and other resource documents from the past and learn about cultures that had significant impact on the evolution and growth of civilization and humanity.

Until now, the education market has been seen as not very profitable and difficult to penetrate, partly due to lacking public investments in the sector. Educational multimedia has, indeed, only just started to enter schools, bringing innovative options for teaching and learning practices and with the Internet opening a number of new possibilities to students. In a generally limited educational climate, edutainment has risen as a much more profitable area by focusing on education as an interactive and entertaining experience and targeting individuals as customers. While schools may still represent interesting future targets for the edutainment sector, these continue to be restricted by specific regional and national polices that are generally unable to keep up with technological advances in the market. On the other hand, families and individuals represent a much more flexible and receptive market for innovative technologies in the edutainment field [Alfaro, 2001].

Edutainment is one product of a major shift - tied to changes in the economy - that is occurring in how we view leisure time in Western societies. In the manufacturing economy, people thought leisure was the reward for hard work. Work was associated with self-improvement and leisure with relaxation that had no other practical use. Today, more people work with their brains than their bodies. People are using their scarce, but more highly valued, leisure time differently, and they have an entirely new attitude to leisure. They see leisure time as an opportunity to improve themselves and their children and do worthwhile things, rather than as purposeless relaxation and entertainment. The change in values was led by changes in the economy, as it moved from manufacturing to technological. Consider the fact that from 1983 to 2000, managerial and professional specialty jobs have increased from 23% to 30% whereas manufacturing employment declined from 16% to 13.5%. Even in manufacturing, many workers now work with their brains, where computerisation and robotics are playing an increasing role in all aspects of manufacturing and distribution. A knowledge society places a high value on education and enrichment.

Lifelong education has become an important priority for many adults. In 1990, 20% of Americans 25 years and older had a bachelor's or higher degree. In 2000, the percentage had increased to 24%, and 52% had attended some college or had a degree. The 1999 National Adult Education Participation Report by the National Center for Education Statistics found that, in 1999, 23% of all adults took one or more college courses strictly for personal enrichment, for non-work related, personal development purposes.

This shift to educational or enriching use of leisure time is also evidenced by the growth in popularity of educational cable television channels such as *The Discovery Channel*, *The History Channel*, *The Travel Channel*, *House & Garden* and *The Learning Channel* [White 2003].

Edutainment is a key focus area for companies. This can be seen, e.g., in the development of commercial games.

One edutainment game, *Versailles 1685, A Game of Intrigue,* takes the player to the magnificent palace of Versailles and the court of King Louis XIV. As the adventure unfolds, you get to explore every corner of the palace and its grounds, find clues, converse with the royal family, the ministers, artists, servants and ladies of the court and unlock the mysteries of the conspiracy that threatens the very survival of Versailles. Another game, *Egypt 1156 BC, Tomb of the Pharaoh,* takes place during the reign of Ramses III at a site that is now called Luxor. You will play Ramose, an ordinary man in extraordinary circumstances, whose father has been accused of arranging the pillage of a royal tomb. [Houston 1999]. *Capital Moves* is an interactive educational game that allows users to match wits with the CEOs of top Canadian corporations, and which has a Mutual Funds Exhibit that uses a touch screen display to test users' basic Mutual Funds expertise, then teaches them more about the Funds market.

The market trend indicates that the school market for occasional-use supplemental software is limited largely to the elementary grades. The home market revenue for high-production-value edutainment software bids fair to overtake revenue for educationally correct, but low-production-value school software, in two to three years. The visibly more sophisticated product may set a new standard of parental expectation for the quality of school software, creating a new demand on the school budget for a higher-priced, educationally correct and visually sophisticated software product. The trend towards image-enriched, visually more sophisticated educational software would be accelerated by the successful entry of "Hollywood-based" firms or subsidiaries like Lucas and Disney into the educational software market.

Edutainment is as much a marketing concept as it is content. Because of edutainment's appeal, more and more entertainment products and venues are marketing themselves as edutainment to increase their perceived value. Even some informal learning institutions are incorporating the word edutainment into their marketing. The New Jersey Aquarium, for example, markets itself as an edutainment experience. Kellogg's Cereal City, a corporate brand museum, is being billed as an edutainment centre.

Edutainment needs to be designed from an understanding of child development and approached "through the eyes of the child," with sensitivity to a child's scale and how they see, interpret and use space and objects. Since much of children's play takes place in their minds through imagination, you need to create the right space (the stage) and supply the right objects (the props) to support their play (act out their scripts). Since children's physical size, skills, and play change as they develop, the edutainment must offer a continuum of challenge, so there is a match between their capabilities and the play opportunities. If there is a mismatch with what children have the interest and ability to do, they will be bored. The consumer edutainment market is seen as far exceeding in size the school market. The industry is currently taking a liberal view of this category, which suggests combining learning with fun. It is argued that most edutainment efforts go wrong because such programs are not tied to schools or curriculum and have no governing standards or links to any form of accreditation [Harvey, 1995].

7.3 Entertainment

Let us now briefly look at the progress of interactive entertainment, partly drawing upon the findings in the CLASS report on Interactive Television [Lazzari 2002].

The Amsterdam office of Forrester Research predicts that sales of goods over interactive TV in Europe will jump from $655 million in year 2002 to $2.1 billion next year. By 2005, it figures that European households with interactive TV will spend $260 a year on items purchased over the TV. Idc reports that in the United States, in 1999, there have been over one million subscriptions for interactive television services. In five years from now this figure is bound to increase progressively, reaching ten million new connections by 2004. According to the Idc survey, this phenomenon will not only affect the United States. Globally, it is forecast that subscriptions will increase from the 6 million recorded last year to19, 5 million in 2004. These figures show that there exists a growing interest in the users towards interactive television services, whatever they may be. Facing this pace of growth, relevant investments could be addressed towards the development of interactive applications based on natural language processing techniques. According to Forrester Research, the ITV market in 2005 should reach, only in the United States, a turnover of 25 billion dollars, of which $ 5bn would be subscriptions, $8 billion commerce, and more than half, $12 billion, would be advertising revenue. It could be assumed that a relevant portion of this turnover could be addressed towards natural language applications.

Microsoft has played an important role in the development of the Internet side of interactive TV. The company has created, since 1997, a sequence of technological agreements and strategic alliances, in order to achieve the goal to provide TV users and Internet surfers the unique box, performing both the role of a PC and that of the digital decoder. For this purpose, Microsoft established a joint venture with Thomson Consumer Electronics, one of the largest manufacturers, for creating an Interactive TV-set. In the meanwhile, Bill Gates signed another agreement with DirectTV and EchoStar, two major operators of satellite TV, to incorporate WebTV Plus, the software to access the network, in the digital receivers. Moreover, Microsoft purchased a 3% share of the AT&T capital. It is a 5 billion dollars investment giving Bill Gates the key to a rich and secure market. The giant of telecommunications will control about 60 per cent of the television cables reaching American households. AT&T acquired Tci first and MediaOne later, the two major Cable-TVs, each with dozens of million subscribers. On the other side, Bill Gates fostered an agreement with a third cable operator, Comcast, a company of which he is a shareholder. If AT&T have the objective of providing telephone services at competitive prices through the television cables, Microsoft objective is to install the boxes carrying high speed Internet on the screen of millions of subscribers to cable television. On the screens, television blended with Internet will offer the same kind of interactivity as a personal computer.

The problem of usability of new technologies is well known to the digital and pay television industry. Also the simplest interactive applications developed by Tps and Stream have been described as "difficult to use" by most subscribers. Making television interactive means adding technology to the existing one and this process could make more difficult to access the new services.

In order to assess all options, it is still important to define the goals wanted to be aimed at, especially when technology is added to technology, as well as to detect the applications that can be translated into a true value added. In other words, new interactive applications must be user friendly, because, today, most technological applications are extremely complicated. "No one ever thinks of making them easy to use and no one ever thinks of normal people who find it impossible to use them" the guru Donald Norman used to say. Moreover, it is necessary to consider that the present offer of technology entertainment is already big and confusing. Each new option will have to bear in mind a reasonable time for the previous offer to "deposit itself" and leave room for the next one.

According to Brian Seth Hurst of Pittard Sullivan, innovations in interactive television must create an experience for users leaving the technological aspect imperceptible. He states also that the areas which will be combined to create an interactive television experience will be not only content, but also navigation and design, i.e., access. If the quality of navigation and design is not high, the right content cannot be individuated and selected. It appears clear that the process towards an interactive television goes through not only the identification of business areas and creation of applications, but, first of all, the creation of tools for making easy, friendly and efficient the identification/selection of content and services.

An ideal solution might be that of introducing the system by means of a virtual presenter, recalling the Talking Head metaphor, used some time ago in a famous Mtv musical programme. The "nice virtual presenter" will not only answer to the user's queries, indicating the choice; above all, it will propose himself a programme selection, on the basis of the preferences of the user he becomes acquainted with better day after day, learning from the choices made by the user and will help the user, step by step, in the programme identification and selection process. Of course, the virtual presenter has his own individual features and character, which can be defined on the basis of the user's choices: i.e., starting up the system, it will be possible to select a "serious" presenter - neutral, professional - or a "friendly" one, capable of playing around with the preferences and options of the user, with ironical word plays, creating entertaining sessions as well. The natural language interface will play a relevant role also in the Video On Demand and Pay Per View services, introducing and commenting the available programmes to be selected.

The application of the interactive presenter, will have the potential of facing other issues, such as, selecting interactive services available on the platform, or as a component of an interactive service, as interactive advertising and t-commerce, or news retrieval or tourist service. A natural language processing application will be part of a more complex interactive business application as the human – system interface, in order to establish an easy and friendly dialogue between user and service. An example might be t-commerce: during a commercial, Talking Head will have a dialogue with the user, providing further information on the product showed in the programme, proposing other products and acquiring the necessary data from the client in case of an interactive purchase transaction.

Natural interactions are a marketing added value for the television industry. The market of digital television is growing to millions and millions of users, either in the pay sector and the terrestrial area, where both free TV and pay services are available. The industry and the platforms are looking for new marketing triggers to help the current commercial offer, to increase users fidelity and revenue per subscriber. Indeed, Britain's BSkyB says it wants to use interactive services to increase its revenue per subscriber from $415 to $585 in the next two years. The number of channels is very high (1500 digital TV channels are operating within EU), the cost of content is increasing year after year (namely for movies and sports), and the typology of content is rich and wide. The competition between the major players is very aggressive and users are looking for quality of services, instead than for quantity of channels. Natural language applications seem to be the right trigger. Interactive personal presentation could be a strong indicator of the level of quality offered by platforms and TV operators.

7.4 Other areas

Customer service is one of the most important aspects of a successful business and, simultaneously, one that requires a great amount of resources. In today's competitive business environment, providing value to the customer is of paramount importance for business to survive. The single most important way to provide value is to let customers feel they have a unique personal relationship with the business. This business model is known as one-to-one marketing [Peppers and Rogers 1997]. Until recently, electronic commerce on the WWW was mass-oriented, business offers and services were uniform and geared to suit the largest possible buyership. Adaptive hypermedia along with natural language generation can provide the basis to establish, maintain and extend a customer base by delivering tailored products and services and thereby create stable long-term relationships with customers.

Significant commercial trends in this area are in-vehicle information and navigation aids for drivers. A number of efforts have already begun from industry to move from canned text to true natural language generation in these applications. The KARMA project, for example, prototyped a system using augmented reality and i3p technology to lead the user on a personalised step-by-step explanation on how to fix a laser printer. The end result can translate not only into a decrease in overall support costs and cost-per-interaction, but also into improved quality of customer support content by better understanding customer requirements [Alfaro 2001].

Automatic text generation, especially when coupled with graphics and other media, has grown into an area of significant promise and importance in the business world. Dynamic document tailoring on the World Wide Web and the use of large existing databases of information as an input to a text generation process, render the field particularly exciting for the commercial market, where demands on output and document generation is of key importance. Trends, particularly in the US, strongly focus on finding commercial applications of natural language generation technology in generating standardised multi-paragraph texts such as business letters or monthly reports.

Training of employees at a company is another area, which can be significantly enhanced by improving learning, such as by presenting personalised information for each user, and by minimizing the internal human resources required to train new employees. Automating business activities such as customer service have a net positive impact, not only on the competitiveness and efficiency of the business, but also in the value added to society. I3p technology has the potential of significantly improving the lives of individuals performing a number of tasks from home by using the telephone or Internet, such as purchasing goods or services, learning interactively, and managing financial accounts. The Internet can provide yet another value to its users, particularly to older people who have difficulties leaving their homes or the handicapped. By improving the possibilities of accessing information, these groups of individuals may stand to profit in ways not traditionally available to them. Technologies such as natural language generation and user modelling can vastly simplify the presentation of information to non-savvy computer users as well as provide services for individuals with various disabilities. In a world of ever-increasing globalisation, the need to provide and receive information in different languages is imperative. With the onset of a unified Europe, the importance of accurate and fast multilingual tools, or tools capable of on-the-spot translations, is a requisite for smooth communication. In Europe, as well as in the US, there is a strong interest on multilingual generation and machine translation and trends indicate that multilingual output will continue to be an important research field [Alfaro 2001].

Acknowledgement

This draft report was supported by the EC Human Language Technologies CLASS (Collaboration in Language and Speech Science Technology) project grant. The support is gratefully acknowledged.

References

[Alfaro 2001] Alfaro Ivana : Intelligent Interactive Information Presentation Field Assessment, ITC-irst, Italy.

[André and Rist 1993] In Maybury, M. T. (Ed.): *Intelligent Multimedia Interfaces.* Cambridge, MA: MIT Press.

[Benoit et al. 2000] Benoit, C., Martin, J. C., Pelachaud, C., Schomaker, L., and Suhm, B.: Audio-visual and multimodal speech systems. In D. Gibbon (Ed.), *Handbook of Standards and Resources for Spoken Language Systems,* Supplement Volume. Dordrecht: Kluwer Academic Publishers.

[Bernsen 1994] Bernsen, N. O.: Foundations of multimodal representations. A taxonomy of representational modalities. *Interacting with Computers* 6, 4, 347-71.

[Bernsen and Dybkjær 2003] Bernsen, N. O. and Dybkjær, L.: Best practice in natural and multimodal interactivity engineering. *CLASS Natural and Multimodal Interactivity Deliverable 1.5 + 1.6.* NISLab, Denmark.

[Bernsen and Stock 2001] Bernsen, N. O. and Stock, O.: *Proceedings of the CLASS Verona Workshop on Intelligent Interactive Information Representation.* ITC-irst, Italy.

[Bernsen et al. 1998] Bernsen, N. O., Dybkjær, H. and Dybkjær, L.: *Designing Interactive Speech Systems*. From First Ideas to User Testing. Springer Verlag.

[Bernsen 2001] Bernsen, N. O.: *Can we "think big" on natural interactivity*. Invited talk, EC Human Language Technologies (HLT), Luxemburg, 13 March 2001. Please contact the author.

[Bernsen et al. 2001] Bernsen, N. O. (Ed.): Speech-related technologies. Where will the field go in 10 years? *ELSNET brainstorming document* v.4, March 2001. See www.elsnet.org/roadmap.html

[Bernsen 2002] Bernsen, N. O.: Multimodality in language and speech systems - from theory to design support tool. To appear in Granström, B. (Ed.): *Multimodality in Language and Speech Systems.* Dordrecht: Kluwer Academic Publishers

[Bolt 1980] Bolt, R. A.: Voice and gesture at the graphics interface. *Computer Graphics,* 262-270.

[De Bra, P 2001] De Bra, P: Course on HyperMedia Structures and systems [online]. Available from Computer Science Department, Eindhoven University of Technology.

[Feiner and McKeown 1993] In Maybury, M. T. (Ed.): *Intelligent Multimedia Interfaces.* Cambridge, MA: MIT Press.

[Harvey 1995] Harvey, J. and Purnell, S.: Workshops on Critical Issues. *The Market for Educational Software.*

[Houston 1999] Houston Tom.: That's Edutainment, June 1999.

[Hovy and Arens 1990] Hovy, E. and Arens, Y.: *When is a picture worth a thousand words* ? Allocation of modalities in multimedia communication. Paper presented at the AAAI Symposium on Human-Computer Interfaces, Stanford.

ISLE project: isle.nis.sdu.dk

LangTech2002: The new European Forum for Language Technology, Berlin.

[Lazzari 2002] Tommaso Maria Lazzari: *Interactive Television: An Overview and Future,* Prospect Report written for the CLASS project.

Meeting recording Project: http://www.icsi.berkeley.edu/~dpwe/research/mtgrcdr/

NICE project: http://www.niceproject.com/

NITE project: nite.nis.sdu.dk

[Peppers and Rogers 1997] Peppers, D. and Rogers, M.: *Enterprise One to One: Tools for Competing in the Interactive Age*, Currency/Doubleday.

[Rist el al. 1997] *Computer Standards and Interfaces,* Special Double Issue on Intelligent Multimedia Presentation Systems, Vol. 18, No. 6-7.

Robita project: tk.elec.waseda.ac.jp/robita/

SmartKom project: smartkom.dfki.de

[Stock and Zancanaro 2002] Stock, O. and Zancanaro, M.: Intelligent Interactive Information Presentation for Cultural Tourism. Proceedings of the International CLASS Workshop on Natural, Intelligent and Effective Interaction in Multimodal Dialogue Systems. Copenhagen, June, 2002, 152-158.

[Turing 1950] Turing, A.: Computing machinery and intelligence. Mind 59, 433-60.

VICO project: www.vico-project.org

[van Kuppevelt et al. 2002] van Kuppevelt, J., Dybkjær, L., and Bernsen, N.O. (Eds.): Proceedings of the CLASS Copenhagen Workshop on Natural, Intelligent and Effective Interaction in Multimodal Dialogue Systems (June 2002).

[Werthner 2001] Werthner Hanne: *Just Business, Shouldn't we have some fun,* e-commerce and Tourism Research lab, irst-ITC and University of Trento, Italy, E-commerce Competence Center (EC3) , Vienna ,Austria.

[White 2003] White, Randy: *That's Edutainment*, White Hutchison Leisure and learning group.