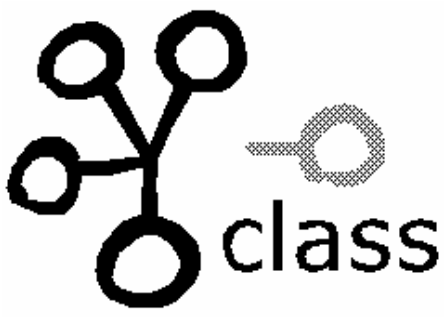

CLASS Natural and Multimodal Interactivity Deliverable D1.5+6

Best practice in natural and multimodal interactivity engineering

February 2003

Authors

Niels Ole Bernsen and Laila Dybkjær
NISLab

Contents

1 Introduction

1.1 Aim of this report

This report aims to map out where we are in natural and multimodal interactivity engineering, where the field is going, and what are the needs which should be met for the field to advance as effectively and efficiently as possible in the years to come. The term 'natural and multimodal interactivity engineering' itself appears to be a new one which may or may not survive this report. What the term refers to, however, is something which has come to stay, namely an ongoing, radical transformation of the field of interactive computer systems.

The specific material on which the report is based are the 21 papers presented by invited speakers and colleagues responding to an open competitive call for papers for the CLASS Workshop on Natural, Intelligent and Effective Interaction in Multimodal Dialogue Systems held in Copenhagen in June 2002. Extended and revised versions of most of the papers are under publication in [van Kuppevelt et al. 2003]. The present report takes a global view of the field of natural and multimodal interactivity engineering and superimposes the contributions of the 21 CLASS papers onto this view to gauge the current state of the field, its needs and future prospects.

We would like to add a note on the methodology for writing this report. The Copenhagen 2002 CLASS workshop was not just (i) a very worthwhile event in an emerging field due to the general quality of the papers presented. It was also (ii) to a large extent representative of the state of the art in the field, as demonstrated by comparison with the scope and contents of topically related workshops and conferences during the last couple of years, cf. Section 1,3. Thirdly (iii), it is impossible with limited authorship to write an early state of the art in a new field by taking everything into account. Analysing in some depth the CLASS workshop papers has provided a focus for our work and a chance to use well-circumscribed evidence to compare and test our own visions for how the field is developing. Finally (iv), as the CLASS workshop papers are now in the book publication pipeline [van Kuppevelt et al. 2003, to appear], we intend to immediately send the present report to the authors of the accepted book chapter contributions and use their feedback to provide an updated version as an introduction to the CLASS book. This version will, we hope, profit from the benefit of having been "vetted" by some of the most prominent current actors in the field.

1.2 Current characteristics of natural and multimodal interactivity engineering

The first and most prominent characteristic of natural and multimodal interactivity engineering is that the field is not yet an established field of research and commercial development but, rather, an emerging one in all respects, including applicable theory, experimental results, platforms and development environments, standards (guidelines, de facto standards, official standards), evaluation paradigms, coherence, ultimate scope, and general topology of the field itself, "killer applications", etc.

A second important characteristic of the field of natural and multimodal interactivity engineering is that its practitioners come from very many different, and often far more established, fields of research and industrial development, such as speech technology, computer graphics, computer vision, human-computer interaction, virtual and augmented reality, non-speech sound, haptic devices, telecommunications, etc. It may be noted that the fact that a field of research has been established over decades in its own right is fully compatible with most of its practitioners being novices in natural and multimodal interactivity engineering. It follows that community formation is an ongoing challenge for all.

Thirdly, the field is expanding very rapidly at present, primarily, it seems, driven by a shared but perhaps not quite unified vision of the potential of new interactive modalities of information representation and exchange for radically transforming the world of computer systems, networks, devices, applications, etc. from the GUI (graphical user interface) paradigm into something which will enable a far deeper and much more intuitive and natural integration of computer systems in people's

lives. The inherent, ultimate vision which the field incorporates is, we believe, one in which even the age-old term "interaction" becomes inadequate for characterising the systems which will emerge. Rather, people will often no longer have to *interact* with the system - i.e. consciously and deliberately inputting information into the system - in order to get things done, the system will do that by itself based on knowledge and observation of its human companions. Thus, with all its dangers and pitfalls, the inherent, ultimate vision of the field might be called *the caring system* with which we will often interact, of course, but which aims to do what we need done whether or not its pursuance of those aims result from traditional human-system interaction.

Fourthly, the field of natural and multimodal interactivity engineering is vast no matter how one looks at it: in terms of the new classes of applications envisioned, research challenges, community integration needs, supporting knowledge and craft skills which can help newcomers off to an early start, future impact, etc. Obviously, the field inherits all or most current trends in today's world of computing more generally, such as ambient intelligence [Ducatel et al. 2001] which, in fact, incorporates most of the others, including ubiquitous computing, higher mobile bandwidth, agent architectures, powerful networks, new devices, etc.

1.3 Multimodality and natural interactivity

Conceptually, natural and multimodal interactivity engineering combines natural interactivity and multimodality. While both concepts have a long history, it would seem that they continue to sit somewhat uneasily side by side in the minds of most of us. *Multimodality* is the idea of being able to choose any input/output modality or combination of input/output modalities for optimising interaction with the application at hand, such as speech input for many heads-up, hands-occupied applications, speech and haptic input and output for applications for the blind, etc. A *modality* is a particular way of representing input or output information in some *physical medium*, such as something touchable, light, sound, or the chemistry for producing olfaction and gustation [Bernsen 2002, see also Carbonell and Kieffer 2002]. The physical medium of the speech modalities, for instance, is sound or acoustics but this medium obviously enables the transmission of information in many modalities other than speech, such as earcons, music, etc. The term multimodality thus refers to any possible combination of elementary or *unimodal* modalities.

Compared to multimodality, the notion of *natural interactivity* appears to be the more focused of the two. This is because natural interactivity comes with a focused vision of the future of interaction with computer systems as well as a relatively well-defined set of modalities required for the vision to become reality. The natural interactivity vision is that of humans communicating with computer systems in the same ways in which humans communicate with one another. Thus, natural interactivity specifically emphasises human-system communication involving the following input/output modalities used in human-human communication: speech, gesture, gaze, facial expression, head and body posture, and object manipulation as integral part of the communication (or dialogue). As the use of touch is marginal in information exchange among humans, the haptic modalities are correspondingly de-emphasised in the natural interactivity vision. Technologically, the natural interactivity vision is currently being pursued vigorously by the emerging research community in animated interface agents as demonstrated, for instance, by the papers presented at the CLASS International Workshop on Information Presentation and Natural Multimodal Dialogue [Bernsen and Stock 2001], the AAMAS (International Joint Conference on Autonomous Agents and Multi-Agent Systems) 2002 conference workshop on embodied conversational agents [Marriot et al. 2002], the PRICAI (Pacific Rim International Conference on Artificial Intelligence) 2002 conference workshop on life-like animated agents tools, affective functions, and applications [http://www.miv.t.u-tokyo.ac.jp/~helmut/pricai02-agents-ws.html], or the HF (Human Factors) 2002 conference workshop on virtual conversational characters: applications, methods, and research challenges [http://www.vhml.org/workshops/HF2002/papers.shtml]. It does, indeed, seem facile to predict that natural and multimodal interactivity engineering will soon need a large conference of its own.

However, a weakness in our current understanding of natural interactivity is that it is not quite clear where to draw the boundary between the natural interactivity modalities and all of those other modalities and modality combinations which could potentially be of benefit to human-system interaction. For instance, isn't pushing a button, although never used in human-human communication

for the simple reason that humans do not have communicative buttons on them, as natural as speaking? If it is, then, perhaps, all or most research for on useful multimodal input/output modality combinations is also research into natural interactivity even if the modalities addressed are not being used in human-human communication? In addition to illustrating the need for more and better theory in the field of natural and multimodal interactivity engineering, the point just made may explain the currently uneasy relationship among the two paradigms of natural interactivity and multimodality. In any case, we have decided to combine the paradigms and address them together as natural and multimodal interactivity engineering.

Finally, by engineering we primarily refer to software engineering. It follows that the somewhat innovative expression natural and multimodal interactivity engineering represents the idea of creating a specialised branch of software engineering for the field addressed in this report.

1.4 A matrix for the field

Roughly speaking, the moving forward of any systems field, such as the field of natural and multimodal interactivity engineering, takes *understanding* of problems and solutions, knowledge and skills for *building* (or developing) the systems, and *evaluation* of any aspect of the process and its results. In the particular case of natural and multimodal interactivity engineering (henceforth NMIE), these goals could be expanded as shown in Table 1.1. Table 1.1 thus aims to specialise, to a modest extent, general software engineering needs for the particular purposes of NMIE.

Using the structure presented in Table 1.1, Table 1.2 shows how the 21 papers contribute to the NMIE field and its needs.

General	**Generic**	**Specific to NMIE**
Understand it	Future visions	Visions, roadmaps, etc., general and per sub-area
	Applicable theory	Applicable theory for any aspect of NMIE
	Empirical work	Controlled experiments, behavioural studies, scenario studies, task analysis on roles of, and collaboration among, specific modalities to achieve various benefits
	Coding and analysis	New quality data resources, coding schemes, coding tools, and standards
Build it	Enabling technologies	New basic technologies needed
	More advanced systems	New, more complex, versatile, and capable system aspects
	Make it easy	Re-usable platforms, components, toolkits, architectures, interface languages, standards, etc.
Evaluate it	All aspects of it	Evaluate components, systems, technologies, processes, etc.

Table 1.1. Needs for progress in natural and multimodal interactivity engineering (NMIE).

A preliminary conclusion based on Table 1.2 is that, for an emerging field which still has not seen any but the simplest of commercial exploitation yet, the NMIE research being done world-wide today already pushes the frontiers in many of the directions needed. In Section 10, we will contrast the picture provided by Table 1.2 by a view of future NMIE needs (Table 10.1).

Specific to NMIE	**Contributions**
Visions, roadmaps, etc., general and per sub-area	Full natural interactive systems building models of their users. The interactive museum of the future.
Applicable theory for any aspect of NMIE	Communication problems.

Controlled experiments, behavioural studies, scenario studies, task analysis on roles of, and collaboration among, specific modalities to achieve various benefits	Effects on communication of animated conversational agents. Spoken input in support of visual search. Gesture and speech for video game playing. Multimodal segmentation of multiple speakers for multi-speaker speech recognition. Animated talking heads for more intelligible and efficient spoken output. Gaze behaviour for more likeable animated interface agents. Audio-visual speech for child language learning. Tutoring robot future scenarios.
New quality data resources, coding schemes, coding tools, and standards	Coding scheme and tool for NMIE systems evaluation. Gesture annotation scheme. standard for internal representation of NMIE data codings. Coding tool for multilevel NMIE data coding.
New basic technologies needed	Interactive robotics: robots controlled multimodally, tutoring robot. Multi-speaker speech recognition. Machine learning: of language. Audio-visual speech synthesis for talking heads.
New, more complex, versatile, and capable system aspects	Multilinguality. Ubiquitous (mobile) application. Location awareness. Web access through spoken dialogue and stylus. Personal assistants. On-line observation-based user modelling for adaptivity. Complex natural interactive dialogue management. Machine learning of language.
Re-usable platforms, components, toolkits, architectures, interface languages, standards, etc.	Platform for mobile systems. Platform for natural interactivity. Re-usable components (many papers). Development toolkit for multimodal dialogue management. Architectures for multimodal dialogue management. Multimodal interface language. VoiceXML. XML for data exchange (many papers).
Evaluate components, systems, technologies, processes, etc.	Framework for multimodal evaluation. Effects on communication of animated conversational agents. Evaluations of talking heads. Evaluation of audio-visual speech synthesis for learning.

Table 1.2. How the natural interactivity and multimodality papers address current NMIE needs.

1.5 Modalities investigated

We have seen in Section 1.3 that multimodality includes all possible modalities for the representation and exchange of information among humans and between humans and computer systems, and that natural interactivity includes an incompletely defined sub-set of those modalities. In view of this open-ended (or wide open) space of unimodal modalities and modality combinations, it may be useful to look at the modalities actually addressed in the workshop papers. These are summarised in the following list:

1. I: speech, pen for pointing; O: speech, text, graphics (maps, POI pictures).
2. I: gesture, speech. O: speech, graphics.

3. I: speech, gesture vs. speech-only. O: embodied conversational agent + images vs. speech-only + images.
4. I: touch gesture, multi-speaker speech. O: speech, images, movement.
5. I: speech. O: graphics.
6. I: text, speech, gesture. O: speech, images.
7. I: speech, gesture. O: speech, gesture.
8. Many (vision paper).
9. I: speech, gesture, object manipulation/manipulative gesture. O: video game.
10. I: speech, camera-based graphics. O: N/A.
11. I: speech, keyboard, mouse, pen-based drawing and pointing, camera. O: graphics and text display.
12. I: speech. O: animated talking head, lip synchrony, visual prosody, gaze for directing user attention, communicative function signalling, internal state signalling, including emotions. Maps, charts, time tables, images.
13. "Technologically mediated communication".
14. I: typed text, O: talking head, gaze.
15. I: gesture. O: N/A.
16. I: N/A. O: talking face.
17. I: speech and gesture from 2nd language learners. O: N/A.
18. I: mouse, keyboard, speech, stylus for pointing gesture, writing, drawing, etc. Focus is on spoken input. O: speech, web graphics.
19. I: speech, haptic buttons. O: music, text, tactile rhythm.
20. I: N/A. O: robot pointing and beat gestures.
21. Many (vision paper).

In the list above, I means input, O means output, and N/A means not discussed in detail, or not relevant. Not surprisingly, the two vision papers [Cole 2002] and [Stock and Zancanaro 2002] mention many different modalities and modality combinations. Combined speech input/output - and this, in fact, means spoken dialogue almost throughout - is addressed in about half of the papers. Two thirds of the papers address gesture input in some form. Six papers address output modalities involving talking heads, embodied animated agents, or robots. Only one paper [Darrell et al. 2002] addresses computer vision input. Emotions expressed through face or gesture is almost absent as a main focus. Despite its richness and key role in natural interactivity, input or output speech prosody is hardly discussed at all. It forms part of the background in [Milde 2002], and, significantly, [Granström and House 2002] discuss graphical ways to *replace* missing output speech prosody by facial expression means.

In general, the unimodal input and output modalities and their combinations discussed at the CLASS workshop would appear to be representative of the state-of-the-art in NMIE. The workshop papers make it quite clear how far we are from mastering the very large number of potentially useful unimodal "compounds" theoretically, in input recognition, in output generation, as well as in understanding and generation.

1.6 Plan for this report

In the following Sections 2 through 9, we briefly review and discuss the NMIE contributions made in the CLASS workshop papers, following the structure of Tables 1.1 and 1.2. Throughout, discussion focuses on what is already being done in the current state-of-the-art to further the global goals of NMIE as well as on what is not being done and what still needs to be done. Based on the discussion, Section 10 presents a view of current NMIE research needs following the now familiar, proposed structure of the field.

2 Visions for natural and multimodal interactivity engineering

Among the CLASS workshop papers, two papers present visions for the NMIE field.

Although brief (a full version will appear in [van Kuppevelt et al. 2003]), the first paper [Cole 2002] highlights several important points. Using natural interactive teaching systems (or tutorial systems) applications for illustration, the paper evidences the important driving role of re-usable platforms, such as the DARPA Communicator system manager or "hub" (http://fofoca.mitre.org), for rapid progress. Moreover, [Cole 2002] points to the future importance to NMIE of two generic technologies which are only modestly represented elsewhere among the papers, see in particular [Darrell et al. 2002]. These technologies are (i) computer vision for processing camera input in order to track, identify, recognise, and interpret users and their communicative behaviours, including lip movements for audio-visual speech recognition, facial expressions, gesture, body posture, object manipulation, the physical situation in which communication takes place, etc., and (ii) on-line observation-based modelling of the behaviour of individual users for subsequent adaptive use of this information. It is only recently that the computer vision community has begun to address issues of natural interactive and multimodal human-system communication, and there is a long way to go before computer vision can parallel speech recognition as a major input medium for NMIE.

On-line observation-based user modelling has a long history in human-computer interaction. In the last 5-10 years, the user modelling field has been converging with NMIE. As witnessed by, e.g., the decade-long series of conferences on user modelling [http://www2.sis.pitt.edu/~um2003/], many research systems now include observation-based user modelling functionality. For the "beyond interaction" vision of caring systems described in Section 1.3, on-line observation-based user modelling is the key which, arguably, will only work properly when supported by machine learning technology (see below).

The second vision paper [Stock and Zancanaro 2002] illustrates a somewhat different type of vision needed in NMIE, i.e. the *application family-specific roadmap* or, rather, in this case, proto-roadmap, since, nowadays, "full" technology roadmaps tend to include timelines for when systems and component technologies can be expected to be ready for commercial development, appear on the market, or otherwise. For an example, see the ELSNET (European Language and Speech Network) time-lined roadmap on challenges for the next ten years of research in speech technologies [Bernsen et al. 2001]. At the time of writing, multiple roadmaps for different NMIE areas are being developed in Europe in response to the European Commission's first calls for proposals for the 6th Framework Programme.

In a final observation on the two vision papers discussed above, the papers aptly illustrate current initiatives in the NMIE field to extend natural and multimodal interaction beyond traditional information systems to new major application areas, such as not only education, which has been around for a while already, notably in the US-dominated paradigm of tutoring systems using animated interface agents, but also to edutainment and entertainment. While the GUI, including the current WWW, might be said to have the edutainment potential of a schoolbook or newspaper, NMIE systems have the much more powerful edutainment potential of brilliant teachers, comedians, and exiting human-human games.

3 The need for applicable theory

It may be characteristic of the NMIE field as a whole at present that our sample of papers only includes a single contribution of a primarily theoretical nature, i.e. [Healey and Thirlwell 2002] which applies a psycholinguistic model of dialogue to help identifying a subset of communication problems in order to judge the effectiveness of multimodal communication. Human-machine communication problems, their nature and identification by human or machine, has recently begun to attract the attention of more than a few NMIE researchers, cf. e.g. the December 2002 ISLE workshop on Dialogue Tagging for Multi-Modal Human Computer Interaction (http://www.research.att.com/~walker/isle-dtag-wrk/), and it has become quite clear that we need a far better understanding of miscommunication in natural and multimodal interaction than we have at present.

Speaking more generally, the fact that we only had a single theoretical paper, if characteristic, is emphatically *not* characteristic in the sense that the field does not need applicable theory. On the contrary, a large number of papers actually do apply existing theory in some form, ranging from empirical generalisations to full-fledged theory of many different kinds. For instance, [Beringer et al. 2002] applies theory on the correlation between cost of communication and user satisfaction, [Bickmore and Cassell 2002] tests generalisations on the effects on communication of involving embodied conversational agents, [Carbonell and Kieffer 2002] applies modality theory, [Chai et al. 2002] applies various theories of human dialogue to the development of a fined-grained semantics-based multimodal dialogue interpretation framework, [Massaro 2002] applies theories of human learning, and [Raggett 2002] builds on existing theory for spoken human-machine dialogue. The point is rather, we submit, that NMIE theory development is hard to do, slows down what we want to accomplish in engineering, tends to be regarded with scepticism by funding agencies, and tends to be received rather thanklessly by fellow researchers because, if the theory is right, they often have to revise their ways of thinking.

As for needs for applicable theory, it is perhaps correct to observe that the NMIE field is unparalleled in its needs for a large variety of theoretical support. All of a sudden, for instance, when developing life-like animated characters, we are moving into high-complexity areas, such as human personality, emotions, character, attitudes, detailed non-verbal behaviour, etc. This suggests that many of the theoretical needs of NMIE can be addressed by seeking to adapt, in order to make them applicable and operational, existing theoretical results from many different disciplines. However, adapting a theory is a theoretical exercise in itself, and it seems likely that we shall need new theory in addition to adaptation of existing theories. Some obvious sources for new theory are the emerging generalisations discussed in Section 4 and the much needed new coding schemes discussed in Section 5.

4 Empirical results

By contrast with theory proper and fortunately so, the NMIE field is replete with empirical studies of human-human and human-machine natural and multimodal interaction. This contrast may be explained by pointing out that, by their nature, empirical studies are much closer to the process of engineering than is theory development. We base the construction of NMIE research systems not only on applied theory but, perhaps to a far greater extent, on hunches, assumptions, extrapolations, untried transfer from different application scenarios, user groups, environments, etc., or even Wizard of Oz studies, see, e.g. [Corradini and Cohen 2002, Bernsen et al. 1998], which are in themselves a form of empirical study. Having built a prototype system, we tend to be keen to find out how far those hunches, etc. got us. Moreover, empirical testing, evaluation, and assessment are integral parts of software and systems engineering, so, all we have to do is to include "hunches testing" in the empirical work on the implemented system which we would be doing anyway.

It may be noted as well that the comparative ease of doing empirical studies as part of the normal business of engineering sometimes tempts us to think that analysis and reporting of empirical experimentation is easy to do and interpret. When this happens, we get published empirical results stating, for instance, that animated speaking interface agents are liked by users. This blatant over-generalisation is then countered by other findings according to which users prefer, e.g., speech-only communication. Given the importance of empirical investigation for NMIE progress, it is perhaps important to emphasise that proper reporting of empirical results is hard to do in its own way, requiring meticulous description of the setup, dependent and independent variables, instructions given to subjects, etc., as well as painstaking analysis to avoid over-generalisation and other forms of misleading presentation of the findings. When proper reporting guidelines are followed, we cannot help meet another inherent characteristic of a large class of empirical studies in the NMIE field. It is that most controlled experimental setups include such a multitude of independent variables that the results obtained are unlikely to generalise much. This point is comprehensively argued and illustrated for the general case of multimodal and natural interactive systems which include speech in [Bernsen 2002]. Still, as we tend to work on the basis of only slightly fortified hunches anyway (cf. above), the results could often serve to inspire fellow researchers to follow them up. Thus, empirical studies are of major importance in guiding NMIE progress.

The primarily empirical workshop papers illustrate well the points made above except for the one on misleading presentation of findings. One cluster of solid findings demonstrate the potential of audio-visual speech output by animated talking heads for child language learning [Massaro 2002] and, more generally, for improving intelligibility and efficiency of human-machine communication, including the substitution of facial animation for the still-missing prosody in current speech synthesis systems [Granström and House 2002]. In counter-point, so to speak, [Darrell et al. 2002] convincingly demonstrates the advantage of using audio-visual *input* for tackling an important next step in speech technology, i.e. the recognition of multi-speaker spoken input. Jointly, these three papers do a magnificent job of justifying the need for natural and multimodal (audio-visual) interaction independently of any psychological or social psychological argument in favour of employing animated conversational agents.

Returning to the unclarified relationship between natural interactivity and multimodality (Section 1.3), a key question seems to be: for which purpose(s), other than harvesting the benefits of using audio-visual speech input/output described above, do we need to accompany spoken human-system dialogue with more or less elaborate animated conversational interface agents? By contrast with spoken output, animated interface agents occupy valuable screen real estate and do not necessarily add information of importance to the users of large classes of applications. Whilst a concise and comprehensive answer to this question is still pending, it seems, [Bickmore and Cassell 2002] goes a long way towards explaining that the introduction of life-like graphical animated interface agents into human-computer spoken dialogue is a tough and demanding proposition. As soon as an agent appears on the display, users tend to switch expectations from talking to a machine to talking to a human. Compared to this finding, the finding in [Heylen et al. 2002] that users tend to appreciate an animated cartoon agent more if it shows a minimum of human-like agent gaze behaviour might, in fact, speak in favour of

preferring cartoon animated agents over life-like animated agents because the former do not run the risk of facing our full set of expectations to human conversational behaviour.

On the multimodal side of the natural interactivity/multimodality semi-divide, several papers address issues of modality collaboration, i.e. how the use of modality combinations could facilitate, or even enable, human-machine interaction tasks that could not be done easily, if at all, using unimodal interaction. [Carbonell and Kieffer 2002] reports on how combined speech and graphics output can facilitate display search, and [Corradini and Cohen 2002] shows how the optional use of different input modalities can improve interaction in a particular virtual environment. Finally, using scenario-based use case analysis, which is, admittedly, a borderline case of empirical investigation which is normally undertaken in the very early stages of systems development, a single paper [Sidner 2002] reaches out towards real conversational interaction. We would like to expand on this last observation.

Despite the fact that the notion of animated conversational interface agents would seem to have reached canonical status in the NMIE community, it remains a fact that real conversational systems are virtually non-existent today. Probably as a result of the fact that most NMIE practitioners come from research fields other than spoken language dialogue systems, one may observe a certain inflation in the use of the notion of conversation. Thus, a conversational system tends to be synonymous with a system which uses speech input-output, even if the system only understands spoken commands in a, say, 50 words vocabulary. Some protagonists of animated conversational interface agents allow that, for the spoken dialogue to qualify as being conversational, the dialogue should be mixed-initiative. However, mixed initiative spoken dialogue is not the same as conversational dialogue. True, mixed initiative dialogue represents a very important step beyond command and control dialogue, use of designer-designed keywords, system-directed dialogue, or user-directed dialogue. But still, today, mixed initiative dialogue remains almost entirely practised in connection with *task-oriented* spoken dialogue. Apart from the envisioning of conversational dialogue in [Sidner 2002] and the arguments for going beyond finite-state representations of dialogue structure in [Clark et al. 2002], no real conversational dialogue is envisioned in the workshop papers. In general, human conversation is *not* task-oriented. Rather, human conversation moves freely among different domains of discourse and rarely seeks to accomplish specific tasks because, if it did, or when is does, it is not conversation any more. As natural interactivity requires conversational dialogue, it would seem preferable to reserve use of the term conversation for describing the real conversational spoken dialogue systems of the future, whether unimodal or multimodal. We may not need much new in terms of basic theory of conversation because there is already such theories available in this area. However, we are likely to need new theory at the design and implementation levels for how to manage the complexity of conversational dialogue which goes far beyond that of task-oriented dialogue.

5 Coding natural interactive and multimodal data

It is perhaps not surprising that we are not very capable of predicting what people will do, or how they will behave, when they interact with computer systems using new modality combinations and possibly also new interactive devices. More surprising, however, is the fact that we are often just as ignorant when faced with predicting natural interactive behaviours which we have the opportunity to observe every day in ourselves and others, such as: which kinds of gestures, if any, do people perform when they are listening to someone else speaking? This example illustrates that, in order to understand the natural interactive ways in which people communicate with one another as well as understanding the ways in which people communicate with the much more limited, current versions of systems for natural and multimodal interaction, we need extensive studies of behavioural data. Even though not all NMIE practitioners may be aware of it, the study of data on natural and multimodal interaction is on its way to becoming a major research area full of potential for new discoveries.

In order to achieve stable and useful results on the behaviours involved in natural and multimodal interaction, we need, first, *high quality data.* A recent report on available natural interactive and multimodal data resources world-wide is [Knudsen et al. 2002b]. First guidelines on how to handle (create, document, etc.) natural interactive and multimodal data resources are presented in [Knudsen et al. 2003]. Second, we need *coding schemes* for all relevant classes of behavioural phenomena involved in natural and multimodal interaction. A recent report on available natural interactive and multimodal coding schemes world-wide is [Knudsen et al 2002a]. The report shows the need for a large variety of new NMIE coding schemes. First guidelines on how to handle (create, document, etc.) natural interactive and multimodal coding schemes are presented in [Dybkjær et al. 2003]. Thirdly, as data coding by hand is a costly and time-consuming process, we need general-purpose *coding tools* which can facilitate the coding and analysis of all or most aspects of natural and multimodal interactive behaviour. A report on available natural interactive and multimodal coding tools world-wide is [Dybkjær et al. 2001]. The report shows that there is no general-purpose coding tool available yet for coding and analysing all or most aspects of natural and multimodal interactive behaviour. The data coding part of NMIE needs even more than what is briefly outlined in this paragraph, such as meta-data standards, but we hope to have demonstrated already the importance of natural and multimodal interaction data coding for progress in NMIE.

As might be expected, a considerable number of workshop papers make use of, or refer to, NMIE data resources, but since none of the papers takes a more principled view on data resource issues, we will not discuss them further here.

Two workshop papers address NMIE needs for new coding schemes. The first of these [Beringer et al. 2002], illustrates the need for data coding in connection with multimodal systems evaluation and presents a coding scheme for this purpose, cf. also Section 9. The second paper [Martell 2002], presents a new, kinematically-based gesture annotation scheme for capturing the kinematic information in gestures from videos of speakers.

Two workshop papers present new NMIE coding tools. The coding tool in [Beringer et al. 2002] was developed to support the use of the multimodal systems evaluation coding scheme mentioned above. This is a typical situation in which NMIE coding scheme creators often find themselves: in the absence of any general-purpose coding tool for NMIE data, the creators have to build the tool themselves. Importantly, [Milde 2002] illustrates the urgency in the field of having general-purpose NMIE coding tools. Although not demonstrably general-purpose yet, the tool called TASX helps move the research frontier in NMIE coding tools forward, especially by enabling some amount of cross-modality coding. Linking the urgent issue of new, more powerful coding tools with the equally important issue of standardisation, [Martell 2002] proposes standards for the internal representation of NMIE codings.

6 Improving enabling technologies

An enabling technology is often developed over a long time by some separate community, such as by the speech recognition community from the 1950s to the late 1980s, and then, when the technology has matured to the point at which practical applications become possible, the technology is transformed into a tool used all over the place, as is the case with speech recognition technology today. NMIE needs a rather large number of enabling technologies and these are currently at very different stages of maturity. Several NMIE enabling technologies, some of which are at a very early stage and some of which are now finding their way into useful applications, are presented in the CLASS workshop papers in the context of application to NMIE problems, including robot interaction and agent technology, multi-speaker interaction and recognition, machine learning, and talking face technology.

[Burke et al. 2002] and [Sidner 2002] both focus on robot interaction. The general domain in both cases is "hosting", i.e. where a virtual or physical agent provides guidance, education, or entertainment based on collaborative goals negotiation and subsequent action in the world. It is clear that a great deal of work remains to be done before robot interaction becomes natural in any approximate sense of the term. For instance, the robot's dialogue capabilities must be strongly improved and so must the embodied appearance and communicative behaviour of the robot. In fact, [Sidner 2002] makes some of the same conclusions as [Bickmore and Cassell 2002], namely that agents need to become far more human-like in all or most respects before they are really appreciated by humans.

[Darrell et al. 2002] address the problem of multi-speaker interaction and of knowing who is addressing the computer when. Their approach is to use a microphone array combined with computer vision. When these two input channels are combined it becomes possible to find out who is talking to the computer. The method is still at a research stage and needs improvements. Another aspect of multi-speaker interaction is found in relation to in-car applications. In-car application developers are faced with the problem of finding out not only when the driver is speaking and when it is rather one of the passengers who is talking, but also when the driver is addressing the system rather than one of the passengers. Some applications use a push-to-activate button to partly overcome the latter problem.

Developers of spoken language applications must cope with the problems resulting from vocabulary and grammar limitations. In spite of having carried out systematic testing, the developer often finds that words are missing when a new user is trying an application. [Dusan and Flanagan 2002] propose machine learning as a way to overcome part of this problem. Using machine learning, the system can learn new words and grammars taught to it by the user in a well-defined way.

[Granström and House 2002] and [Massaro 2002] both describe the gain in intelligibility that can be obtained by combining speech synthesis with a talking face. There is still much work to do both as regards synthesis and as regards face articulation. Speech synthesis is for most languages still not very natural to listen to and if one wants to develop a particular voice which has to fit a certain animated character, this is not immediately possible with today's available technology. With respect to face articulation, faces need to become much more natural in terms of, e.g., gaze, eyebrow movements, lip and mouth movements, and head movements, as this seems to influence users' perception of the interaction [Granström and House 2002, Heylen et al. 2002].

One of the important enabling technologies for NMIE which is not mentioned in the CLASS workshop papers is audio-visual speech recognition. Not only humans' perception of speech is improved when visual cues are added. The same seems to be true for computers. Thus, audio-visual speech recognition is seen as a way in which speech recognition may be improved. Another important enabling technology is prosody recognition. Work on prosody recognition has been going on for decades at what appears to be a rather slow pace, and we are still far from being able to harness the technology for NMIE purposes. Many cues in the speech input signal will continue to be lost as long as recognisers cannot cope with prosody, which again means important losses in the naturalness of the system's dialogue behaviour. Computer vision and multi-party speech recognition are two other enabling technologies for NMIE which need further progress.

7 Building more advanced systems

Enabling technologies for NMIE are often component technologies and their description, including state of the art, currently addressed research challenges, and unsolved problems, can normally be done in a relatively systematic and focused manner. By contrast with enabling technologies for NMIE which may or may not be applied to problems in this field in the process of reaching maturity, it is far more difficult to systematically describe the complexity involved in the constant push in research and industry for exploring and exploiting new NMIE application types and new application domains, addressing new user populations, increasing the capabilities and sophistication of systems in familiar domains of application, exploring known technologies with new kinds of devices, etc. At the level of description of this report, the picture is one of pushing present boundaries in most directions. Not in all directions, however, because it is still possible to spot surprisingly underdeveloped areas of research and industrial development, i.e. areas which, for instance, one would have thought, the enabling technologies are in place and the user (or consumer) interest is strong, but where amazingly little is happening nevertheless. During the last few years, a core trend in NMIE has been to combine different modalities in order to build more complex, versatile and capable systems, and to get closer to natural interactivity than what is possible with only a single modality involved. This trend is reflected in several of the CLASS workshop papers.

Part of the NMIE paradigm is that systems must be available whenever and wherever convenient and useful, cf. Section 1.2. Thus, ubiquitous computing has become an important application domain. Ubiquitous computing may, e.g., be enabled through wiring up one's home but it may also be enabled via mobile applications. Mobile devices such as telephones, PDAs and portable computers of any (portable) size have become very popular and are gaining more and more functionality. The tourist guide application described in [Almeida et al. 2002] is an example of an application which is meant for mobile use and which runs on an iPAQ pocket PC. Location-awareness may be added to mobile applications in order to augment the mobility dividend. If a system is equipped with GPS, the system knows the user's geographical position. This makes it possible to handle spoken user input, such as "where am I now" or "are there any restaurants in the neighbourhood" in a meaningful way.

Web-applications are gaining ground and many mobile devices offer Internet access. A web page is no longer equivalent to a plain html page which can just be viewed in a browser. Languages like VoiceXML and SALT [Raggett 2002], see also Section 8, support voice-based access to web pages.

It may be difficult for users to know how to interact with new sophisticated applications. [Almeida et al. 2002] concluded that their users needed instructions before they could benefit from the (spoken multimodal) application described. Better built-in user guidance would seem to be needed in this case. However, there are additional ways in which to support a user during interaction. A couple of papers mention user modelling [Chai et al. 2002, Cole 2002]. User modelling may be done in different ways and may happen via information acquired off-line or during interaction. In the latter case, the user model may be updated on the fly or only between interactions. In any case, the idea is to enable a more natural and adequate interaction based on the system's knowledge about the user's preferences, habits, etc. Machine learning is another way in which to increase interaction support. [Dusan and Flanagan 2002] propose to increase the vocabulary and grammar of a system by letting users teach the system new words and their meaning and use.

Increasingly advanced systems also require increasingly advanced and complex dialogue management, cf. [Chai et al. 2002, Clark et al. 2002]. As discussed in Section 4, conversational dialogue has become a buzzword although it would be an exaggeration to call any existing spoken or spoken-cum-animated dialogue system conversational. Real conversational dialogue is hinted at in [Sidner 2002] but the term is being used by many outside the CLASS workshop papers to denote applications which in fact are far from being conversational. Just like real life-likeness of animated interface agents, real conversational dialogue is among the key challenges to be overcome before we can achieve the NMIE vision.

The multilinguality of systems is an important goal which is not merely one of adding speech and language processing for different languages to applications. Multilinguality research also needs to overcome unsolved issues, such as language recognition, user modelling of the user's preferred

language, the enormous challenge of recognising cross-language pronunciation variants, distributed speech recognition for limited-power devices, etc., which are beyond the scope of this report. Multilingual applications are addressed in [Almeida et al. 2002, Reithinger et al. 2002]. In both cases, the application is running on a handheld device.

Multi-speaker input speech is mentioned by [Burke et al. 2002, Chai et al. 2002, Darrell et al. 2002]. For good reason, recognition of multi-speaker input is becoming a hot research topic. We need solutions in order to, e.g., build meeting minute-takers, separate the focal speaker's input from that of other speakers, exploit the potential of spoken multi-user applications, etc.

8 Building systems easily

Due to the complexity of multimodal natural interaction it is becoming dramatically important to be able to build systems easily. It seems likely that no single research lab or development team in industry, including giants such as Microsoft, is able to master all of the enabling technologies required for NMIE progress. Therefore, to advance efficiently, everybody needs easy access to the components and their built-in know-how which are not in development focus. This, again, implies strongly increased attention to issues, such as re-usable platforms, components and architectures, development toolkits, interface languages, data formats, and standardisation.

The DARPA Communicator hub (http://fofoca.mitre.org/) which supports a modular plug and play approach has been used as the underlying inter-module communication framework of the system reported in [Almeida et al. 2002]. Both [Cole 2002] and [Burke et al. 2002] are considering the DARPA Communicator as a candidate for their future work. [Burke et al. 2002] have used the Open Agent Architecture (OAA) (http://www.ai.sri.com/~oaa/) which is a framework for integrating a community of heterogeneous software agents in a distributed environment. Another solution for inter-module communication mentioned by [Burke et al. 2002] is CORBA (Common Object Request Broker Architecture) (http://www.corba.org/). CORBA is not used by any of the reported systems at the workshop but is, e.g., being used in the VICO project on spontaneous speech, mixed initiative in-car information access (http://www.vico-project.org). In any case, what these and other architectural frameworks aim to do is to provide a means for modularisation, synchronous and asynchronous communication, well-defined inter-module communication via some interface language, such as IDL (CORBA) or ICL (OAA), and the possibility of implementation in a distributed environment.

XML (Extensible Markup Language) which is a simple, flexible text format derived from SGML (ISO 8879), is becoming quite popular as, among other things, a message exchange format. It is used for that purpose by, e.g., [Almeida et al. 2002, Burke et al. 2002, Reithinger et al. 2002]. Using XML for wrapping messages to be sent between modules is one way to overcome the problem of different programming languages being used for implementing different modules. XML is becoming a standard for data exchange not only internally between system components but also, e.g., for annotation files, cf. [Milde 2002]. XML is one of the activities in which W3C (the World Wide Web Consortium) (http://www.w3.org/) is involved. Other W3C activities include, as also mentioned by [Raggett 2002]: VoiceXML which is a markup language designed for creating spoken dialogues; a multimodal interaction activity which seeks to develop markup specifications for synchronisation across multiple modalities and devices with a wide range of capabilities; SMIL (Synchronized Multimedia Integration Language) which enables simple and time-controlled authoring of interactive audiovisual presentations, and the W3C 2002 Multimodal Interaction Activity: http://www.w3.org/2002/mmi/. SALT (Speech Application Language Tags) is an alternative proposal to Voice XML founded by Microsoft, among others. SALT is a set of extensions to existing markup languages, in particular HTML and XHTML, that enable multimodal and telephony access to information, applications and web services.

Some of the workshop papers express a need for reusable components. In many cases, the applications described use off-the-shelf software. This is in particular true for mature enabling technologies, such as speech recognition and synthesis components. As regards multimodal dialogue management, there is also an expressed need for reuse in, e.g., [Burke et al. 2002 and Clark et al. 2002]. They have actually tried to reuse components for dialogue management but experienced various problems symptomatic of reuse attempts in an expanding field, such as that of NMIE.

In conclusion, there are already architectures, platforms, and software components available which facilitate the easy building of new NMIE applications, and standards are underway for certain aspects. There is still much work to be done on standardisation, new and better platforms, and improvement of component software. However, in addition we need, in particular, more and better toolkits in support of system development and we need a better understanding of those components which cannot be bought off-the-shelf and which typically are difficult to reuse, such as dialogue managers. Advancements such as these are likely to require significant corpus work. Corpora with tools and annotation schemes as described by [Martell 2002] are exactly what is needed in this context.

9 Evaluation

Software systems and components evaluation is a broad area ranging from technical evaluation over usability evaluation to customer evaluation. Customer evaluation has never been a key issue in research but has rather tended to be left to the marketing departments of companies. Technical evaluation and usability evaluation, including evaluation of functionality from both perspectives, are, on the other hand, considered important research issues.

The papers from the CLASS workshop show a clear trend to focus on the usability aspects of evaluation and on comparative performance evaluation.

Comparative performance evaluations objectively compare users' performance on different systems with respect to, e.g., how well they understand speech-only versus speech combined with a talking face or with an embodied animated agent [Granström and House 2002, Massaro 2002, Bickmore and Cassell 2002]. The usability issues evaluated all relate to users' perception of a particular system and include parameters, such as life-likeness, credibility, reliability, efficiency, personality, ease of use, and understanding quality [Heylen et al. 2002, Bickmore and Cassell 2002, Beringer et al. 2002].

It is hardly surprising that performance evaluation and usability issues are considered key topics today. We know little about what happens when we move towards increasingly multimodal and natural interactive systems which include multiple modalities, both as regards how these new systems will perform compared to alternative solutions and as regards how the systems will be received and perceived by their users. We don't have methods today which enable prediction of how well users will receive a system. We just know that a technically optimal system is not enough to produce user satisfaction.

Two papers from the workshop address how intelligibility of what is being said can be increased through visual articulation [Granström and House 2002, Massaro 2002]. [Granström and House 2002] have used a talking head in several applications, including tourist information, real estate (apartment) search, aid for the hearing impaired, education, and infotainment. Evaluation has shown a significant gain in intelligibility for the hearing impaired when a talking face is added. Eyebrow and head movements enhance perception of emphasis and syllable prominence. Over-articulation may be useful as well when there are special needs for intelligibility. The findings in [Massaro 2002] support these promising conclusions. His focus has been on applications to the hard-of-hearing, children with autism, and child language learning more generally.

[Granström and House 2002] also address the increase in efficiency of communication/interaction produced by using an animated talking head for system output. Probably, naturalness is a key point here. This is suggested by the findings in [Heylen et al. 2002] who made controlled experiments on the effects of different eye gaze behaviours of a cartoon-like talking face on the quality of human-agent dialogues. The most human-like agent gaze behaviour led to higher appreciation of the agent and more efficient task performance.

[Bickmore and Cassell 2002] evaluate the effects on communication of an embodied conversational real-estate agent versus an over-the-phone version of the same system, cf. [Cassell 2000]. Users liked the system better in the speech-only condition. In the embodied condition, they wanted to get down to business. The implication is that the physical embodiment has strong effects on the interlocutors. Users tend to compare their animated agent interlocutors with humans rather than machines. To work with users, animated agents need considerably more naturalness and personally attractive features communicated non-verbally. This imposes a tall research agenda on both speech and non-verbal output, requiring conversational abilities both verbally and non-verbally.

[Beringer et al. 2002] address usability evaluation of multimodal systems from a general point of view. They are inspired by what is perhaps the most widely used framework for usability evaluation of spoken language dialogue systems called PARADISE (Paradigm for Dialogue System Evaluation) [Walker1997]. This framework tries to model user satisfaction as a function of task success and various dialogue cost metrics. The PARADISE model is not unproblematic in itself [Dybkjær et al. 2003] and, as pointed out by [Beringer et al. 2002], the model will have to be extended for multimodal systems evaluation. Based on PARADISE, [Beringer et al. 2002] therefore propose an evaluation

framework which extends the PARADISE approach called PROMISE (Procedure for Multimodal Interactive System Evaluation), for multimodal dialogue systems evaluation. A graphical evaluation tool has been developed which allows the evaluator to compare subjective usability scores from a questionnaire with corresponding objective measures, supporting the practical use of PROMISE. This is relatively early work.

Jointly, the workshop papers on evaluation demonstrate a broad need for performance evaluation, comparative as well as non-comparative, that can inform us on the possible benefits and shortcomings of new natural interactive and multimodal systems. The papers show a similar need for usability evaluation that can help us find out how users perceive these new systems, and a need for finding ways in which usability and user satisfaction might be correlated with technical aspects in order for the former to be derived from the latter.

The workshop papers do not really address technical evaluation other than comparative performance evaluation of certain parameters, and only one paper addresses evaluation of mobile systems and systems on small devices [Almeida et al. 2002]. Several ongoing research projects have as part of their agenda to look into evaluation methods for various aspects of natural interactive and multimodal dialogue systems, for instance: SMADA, Speech Driven Multimodal Automatic Directory Assistance, 2000-2002, http://smada.research.kpn.com/MainPage/, looked at usability issues for small terminals for mobile internet access, cf. [Almeida et al. 2002]; INSPIRE, Infotainment management with speech interaction via remote-microphones and telephone interfaces, 2002-2004, http://www.inspire-project.org/, looks at usability and acceptability evaluation; MIAMM, Multidimensional Information Access using Multiple Modalities, 2001-2004, http://www.loria.fr/projets/MIAMM, cf. [Reithinger et al. 2002] looks at evaluation methods and protocols for multimodal interaction; and NICE, Natural Interactive Communication for Edutainment, 2002-2005, http://www.niceproject.com/, looks at new ways of evaluating natural human-system interaction. However, it would seem timely to establish a high-profile, community-wide project which could address best practice in evaluation for natural and multimodal interactivity engineering at a more overall level, perhaps along lines similar to what was done in the DISC project (www.disc2.dk) which focused on spoken dialogue systems engineering, and including aspects of usability evaluation inspired by the approach taken in the DARPA communicator project which also addressed spoken dialogue systems.

10 Future needs of natural and multimodal interactivity engineering

Based on the discussion of the workshop contributions in Sections 2 through 9 above, Table 10.1 summarises our conclusions about important research needs and working technologies for NMIE. NMIE is an enormous area, and Table 10.1 is no doubt inadequate in several ways. For one thing, the table if a rather global or coarse-grained one as befits the wideness of the field addressed in it. It is easily possibly to zoom in on any particular entry and expand it by providing additional detail. That is why we need to collect existing sub-area-specific roadmaps and create the ones which turn out to be missing. In the speech and language area, for instance, the ELSNET roadmap provides far more detail than does Table 10.1 [Bernsen et al. 2001]. Secondly, Table 10.1 is no doubt to some extent partial through being tainted by the authors' own outlook upon NMIE. Briefly, their background is one of solid experience in spoken dialogue systems and growing experience in animated interface agent interaction, considerable experience in the data resources, coding scheme, and coding tool sub-area of NMIE, considerable experience in all manner of evaluation, good knowledge of modality theory, and extensive experience in technology forecasting, brainstorming and roadmapping, but less solid hands-on experience with computer vision and input/output haptics.

10.1 A matrix view

Specific to NMIE	**Current NMIE needs**
Visions, roadmaps, etc., general and per sub-area	Roadmapping has become popular: we need to collect and integrate them, timeline them, and do those which are missing.
Applicable theory for any aspect of NMIE	Encourage and support development of more applicable theory on NMIE interaction, systems, components, and evaluation.
Controlled experiments, behavioural studies, scenario studies, task analysis on: roles of, and collaboration among, specific modalities to achieve various benefits	More of the same: As many empirical results as we can get. Investigations of new modality combinations for new users, new environments, new applications, etc. Arguably, increase awareness in the field of best practice in conducting, analysing and reporting empirical findings.
New quality data resources, coding schemes, coding tools, and standards	Much more high-quality NMIE data resources, well-documented, re-usable, easy to find on the web, free for research purposes, based on standards. Many more consolidated NMIE coding schemes for new areas, created and documented according to standards. General-purpose coding tools for multilevel, cross-level and cross-modality NMIE data coding. Stronger awareness among practitioners of the area's importance.
New basic technologies needed	Major advances in computer vision for processing camera input in order to track, identify, recognise, and interpret users and their communicative behaviours, including lip movements for audio-visual speech recognition, facial expressions, gesture, body posture, object manipulation, the physical situation in which communication takes place, emotions, personality, etc. Prosody recognition. Multi-speaker speech recognition. Audio-visual speech recognition.

	Prosody synthesis. Machine learning for adaptation, language learning, etc.
New, more complex, versatile, and capable system aspects	Applications for small mobile devices. Situation awareness. Easy production of flexible human-like graphical interface agent behaviours. Real conversational spoken and multimodal dialogue systems (domain-oriented systems). Edutainment and entertainment applications.
Re-usable platforms, components, toolkits, architectures, interface languages, standards, etc.	Easy transfer of all relevant NMIE progress to web applications. Freeware, open source, and other versatile plug-and-play platforms for NMIE. More re-usable components, component interface standardisation. Freeware, open source, and other development toolkits for NMIE. Extension of existing architectures to full NMIE capability, including, e.g., stable generic models for input fusion and output fission.
Evaluate components, systems, technologies, processes, etc.	More knowledge on the usability of different modality combinations. More knowledge on the parameters behind user satisfaction to enable better prediction of user satisfaction. Better methods for usability evaluation. Technical evaluation parameters for natural interactive and multimodal systems.

Table 10.1. Current research needs for natural and multimodal interactivity engineering.

10.2 Discussion

Let us now present some comments on Table 10.1.

10.2.1 Visions, roadmaps, etc., general and per sub-area

See Section 10.1.

10.2.2 Applicable theory for any aspect of NMIE

We strongly feel that this topic, highly neglected as it tends to be, merits a separate report which should be produced in consultation with the NMIE community. A general comment is that theory development is very different from any other NMIE exercise proposed in this report. Theory development is somewhat remote from, or runs in a different time from, most NMIE activities. Theory development is also difficult to fund for the simple reason that it is often very hard to determine which theory development effort, and which theory developer, holds the promise of making a major contribution to the field. Moreover, theory development is often being done by a single individual. In the present climate in Europe, at least, of collaborative transnational project funding, single individuals are prevented even by basic funding rules to have their work supported for some length of time, however important this work might be to a wider community.

10.2.3 Controlled experiments, behavioural studies, scenario studies, task analysis on roles of, and collaboration among, specific modalities to achieve various benefits

Everybody does these things from time to time and given the hunch-based nature of many NMIE research activities (cf. Section 4), it is probably best, and not just simplest, to recommend much more of the same. Strong awareness of the limited extent to which experimental results generalise is important.

10.2.4 New quality data resources, coding schemes, coding tools, and standards

We believe to have stated the case above. The TASX coding tool presented at the CLASS workshop represents interesting progress. We refer the reader to the ISLE report series for more information on other efforts towards building general purpose NMIE coding tools, such as NITE (nite.nis.sdu.dk), Atlas (http://www.nist.gov/speech/atlas/index.html), Anvil (www.dfki.de/~kipp/anvil), etc.

10.2.5 New basic technologies

In this section of Table 10.1, we list a series of enabling NMIE technologies most of which are noticeably absent in the workshop papers.

Given the current state of machine vision and the very difficult problems which still remain to find practical working solutions, and given the crucial role of machine vision for processing all human communicative input behaviour (listed in Table 10.1) which is basically conveyed through the physical medium of light, machine vision needs all the support it can get in order to begin to deliver solutions whose practical importance could equal that of speech recognition. We simply do not know how strongly to recommend substitutes for visual systems, such as data gloves, data suits, etc. For most practical purposes of interaction, these substitutes are probably unwieldy and it would also hardly feel natural for a user to have to dress up in this special equipment.

Speaking about speech recognition, it is sometimes ignored that the field still has to solve several major problems, including that of practically useful multi-speaker recognition and audio-visual speech recognition. In view of the importance of speech in natural and multimodal interaction as evidenced by the workshop papers, solution to those problems must come high on any list of NMIE research priorities. Almost completely absent in the workshop papers is mention of the problem of speech prosody recognition. If we want future NMIE systems to recognise, rather than to loose, information crucial to semantic understanding, speech act understanding, emotion, attitude, personality, etc. interpretation, it is necessary to solve the prosody recognition problem.

As for speech synthesis, several of the workshop papers have presented significant progress in audio-visual speech synthesis, which is why is does not appear in Table 10.1. Output speech prosody, however, remains a major problem. For many languages, it is today possible to choose from a selection of speech synthesisers. However, there is a rapidly increasing need for the spoken output of NMIE systems to reflect contextually determined emotions, personality, attitude, etc. If that need is not met by better prosody synthesis, we will soon have strongly non-verbally expressive talking heads and embodied interface agents which are unable to produce the accompanying speech output modulation.

Finally, as the workshop paper on language learning intriguingly demonstrates, there is a very important role for machine learning in future NMIE systems. Despite progress in making platforms, components, interfaces, architectures available to advanced research, NMIE systems are becoming so complex to build that, ultimately, machine learning would seem to constitute the only way out. This statement may well apply to most of the functionalities needed to achieve full natural interactive systems, including input recognition, input understanding, knowledge base and library building, output generation, and output display/synthesis.

10.2.6 New, more complex, versatile, and capable system aspects

In Table 10.1, this NMIE aspect sits uneasily next to the list of emerging technologies discussed above. It was useful for describing, in Table 1.2, some of the more detailed system innovations presented in the workshop papers. However, as regards future needs, it is an enormous task to venture into describing the many different families of applications which will be enabled by NMIE, and, in terms of pin-pointing real innovation needs, the outcome would not be in proportion to the effort. As regards families, or perhaps rather large minorities, of applications, we only mention the potentially very large and, so far, essentially untapped, areas of edutainment and entertainment applications. We are actually surprised that these large minorities among future NMIE applications were only mentioned, and not described in detail at all, in a couple of workshop papers, especially since we believe that this lack of attention to edutainment and entertainment applications is typical of the NMIE field at present.

We mention three other generic needs under the present heading. The first one is *situation awareness* which is absent in the workshop papers. With situation awareness, the system receives, processes, and makes use of information about the larger system of which it forms part, such as information about the human carrying the (mobile) system, the car in which it is installed, the virtual world concerning which the system enables interaction, etc. The second, similarly absent, need is that of *facilitating the generation of flexible non-verbal graphical natural and multimodal output.* Given enough time and resources, virtually anything is possible in the NMIE field. The current state of the art in natural interactivity output graphics may be compared to the use of concatenated speech for speech synthesis, which is being done by recording relevant output sentences produced by a specific human individual, cutting the input into pieces, and re-assembling the pieces according to a concatenation program at run-time. The output quality is excellent but the flexibility is zero unless it is possible to persuade the human involved in the recordings to re-record new sentences, phrases and words. To efficiently produce the NMIE systems of the future, we need system capabilities corresponding to existing methods for full speech synthesis.

A third need which is, perhaps, more remarkably absent in the workshop papers, is the requirement of a breakthrough from *task-oriented* spoken dialogue systems to *domain-oriented spoken dialogue systems.* Task-oriented spoken dialogue systems, architectures, components, etc. are well described in the literature already, including their mixed initiative and spontaneous user input speech varieties. Surprisingly, as mentioned already (Section 4), such system are standardly described as conversational systems. However, conversation is not task-oriented. Rather, conversation may be characterised by the fact that *no* specific task is being addressed by the interlocutors. Rather, the partners in conversational dialogue move freely among a set of often hard-to-delimit *topics* in a *domain* of discourse, and they even move freely among different domains. So, the next step in spoken dialogue systems beyond the paradigm of task orientation, must be the domain-oriented spoken dialogue system. Evidently, this step requires some form of breakthrough in dialogue management for which the abandonment of hard-coded finite state dialogue structures is only a start. Not least in view of the potential of NMIE edutainment and entertainment applications, domain-oriented conversational systems would be high on the list of future NMIE needs.

10.2.7 Re-usable platforms, components, toolkits, architectures, interface languages, standards, etc.

There are architectures, platforms, and component software available in support of the easy building of new applications. Some are free and others must be paid for. There is also some standardisation work going on. But there is definitely room for much more work both on standardisation and on all kinds of supporting software, be it architectures and platforms, system components, or development tools, including tools for corpus work as mentioned in Section 5.

10.2.8 Evaluate components, systems, technologies, processes, etc

Currently, there seems to be a major interest in usability evaluation aspects and in comparing the performance of new natural interactive and multimodal systems to, e.g., other, often unimodal, systems. For every application we build, we need a great deal of evaluation and that need is not likely to go away. However, more knowledge on how and what to evaluate and, not least, more knowledge on what triggers certain user reactions to software and what are the parameters which influence user satisfaction is highly needed.

11 References

11.1 Papers

Almeida, L., Amdal, I., Beires, N., Boualem, M., Boves, L., den Os, L., Filoche, P., Gomes, R., Knudsen, J. E., Kvale, K., Rugelbak, J., Tallec, C. and Warakagoda, N.: Implementing and Evaluating a Multimodal Tourist Guide. Proceedings of the International CLASS Workshop on Natural, Intelligent and Effective Interaction in Multimodal Dialogue Systems. Copenhagen, June, 2002, 1-7.

Beringer, N., Hans, S., Louka, K. and Tang, J.: How to Relate User Satisfaction and System Performance in Multimodal Dialogue Systems? - A Graphical Approach. Proceedings of the International CLASS Workshop on Natural, Intelligent and Effective Interaction in Multimodal Dialogue Systems. Copenhagen, June, 2002, 8-14.

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

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

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

Bickmore, T. and Cassell, J.: Phone vs. Face-to-Face with Virtual Persons. Proceedings of the International CLASS Workshop on Natural, Intelligent and Effective Interaction in Multimodal Dialogue Systems. Copenhagen, June, 2002, 15-22.

Burke, C., Harper, L. and Loehr, D.: A Dialogue Architecture for Multimodal Control of Robots. Proceedings of the International CLASS Workshop on Natural, Intelligent and Effective Interaction in Multimodal Dialogue Systems. Copenhagen, June, 2002, 23-26.

Carbonell, N. and Kieffer, S.: Do Oral Messages Help Visual Exploration? Proceedings of the International CLASS Workshop on Natural, Intelligent and Effective Interaction in Multimodal Dialogue Systems. Copenhagen, June, 2002, 27-36.

Chai, J., Pan, S. and Zhou, M. X.: MIND: A Semantics-based Multimodal Interpretation Framework for Conversational Systems. Proceedings of the International CLASS Workshop on Natural, Intelligent and Effective Interaction in Multimodal Dialogue Systems. Copenhagen, June, 2002, 37-46.

Clark, B., Bratt, E. O., Peters, S., Pon-Barry, H., Thomsen-Gray, Z. and Treeratpituk, P.: A General Purpose Architecture for Intelligent Tutoring Systems. Proceedings of the International CLASS Workshop on Natural, Intelligent and Effective Interaction in Multimodal Dialogue Systems. Copenhagen, June, 2002, 47-50.

Cole, R.: Perceptive Animated Interfaces: The Next Generation of Interactive Learning Tools. Proceedings of the International CLASS Workshop on Natural, Intelligent and Effective Interaction in Multimodal Dialogue Systems. Copenhagen, June, 2002, 51.

Corradini, A. and Cohen, P. R.: On the Relationships Among Speech, Gestures, and Object Manipulation in Virtual Environments: Initial Evidence. Proceedings of the International CLASS Workshop on Natural, Intelligent and Effective Interaction in Multimodal Dialogue Systems. Copenhagen, June, 2002, 52-61.

Darrell, T., Fisher, J. and Wilson, K.: Geometric and Statistical Approaches to Audiovisual Segmentation for Untethered Interaction. Proceedings of the International CLASS Workshop on Natural, Intelligent and Effective Interaction in Multimodal Dialogue Systems. Copenhagen, June, 2002, 62-71.

Ducatel, K., Bogdanowicz, M., Scapolo, F., Leijten, J. and Burgelman, J.-C.: Scenarios for Ambient Intelligence in 2010. Draft Final Report Version 2. IPTS, Seville, Spain.

Dusan, S. and Flanagan, J.: An Adaptive Dialogue System Using Multimodal Language Acquisition. Proceedings of the International CLASS Workshop on Natural, Intelligent and Effective Interaction in Multimodal Dialogue Systems. Copenhagen, June, 2002, 72-75.

Dybkjær, L., Berman, S., Kipp, M., Olsen, M. W., Pirrelli, V., Reithinger, N. and Soria, C.: Survey of Existing Tools, Standards and User Needs for Annotation of Natural Interaction and Multimodal Data. ISLE Deliverable D11.1, January 2001.

Dybkjær, L., Bernsen, N. O., Knudsen, M. W., Llisterri, J., Machuca, M., Martin, J.-C., Pelachaud, C., Riera, M. and Wittenburg, P.: Guidelines for the Creation of NIMM Annotation Schemes. ISLE Deliverable D9.2, February 2003.

Dybkjær, L., Bernsen, N. O. and Minker, W.: Overview of Evaluation and Usability. In Minker, W., Buhler, D. and Dybkjær, L. (Eds.): Spoken Multimodal Human-Computer Dialogue in Mobile Environments. Kluwer Academic Publishers, 2003 (to appear).

Granström, B. and House, D.: Effective Interaction with Talking Animated Agents in Dialogue Systems. Proceedings of the International CLASS Workshop on Natural, Intelligent and Effective Interaction in Multimodal Dialogue Systems. Copenhagen, June, 2002, 76-82.

Healey, P. G. T. and Thirlwell, M.: Analysing Multi-Modal Communication: Repair-Based Measures of Communicative Co-ordination. Proceedings of the International CLASS Workshop on Natural, Intelligent and Effective Interaction in Multimodal Dialogue Systems. Copenhagen, June, 2002, 83-92.

Heylen, D., van Es, I., Nijholt, A. and van Dijk, B.: Experimenting with the Gaze of a Conversational Agent. Proceedings of the International CLASS Workshop on Natural, Intelligent and Effective Interaction in Multimodal Dialogue Systems. Copenhagen, June, 2002, 93-100.

Knudsen, M. W., Bernsen, N. O., Dybkjær, L., Hansen, T., Mapelli, V., Martin, J.-C., Paulsson, N., Pelachaud, C., and Wittenburg, P.: Guidelines for the Creation of NIMM Data Resources. ISLE Deliverable D8.2, February 2003.

Knudsen, M. W., Martin, J.-C., Dybkjær, L., Ayuso, M. J. M, N., Bernsen, N. O., Carletta, J., Kita, S., Heid, U., Llisterri, J., Pelachaud, C., Poggi, I., Reithinger, N., van ElsWijk, G. and Wittenburg, P.: Survey of Multimodal Annotation Schemes and Best Practice. ISLE Deliverable D9.1, 2002a.

Knudsen, M. W., Martin, J.-C., Dybkjær, L., Berman, S., Bernsen, N. O., Choukri, K., Heid, U., Mapelli, V., Pelachaud, C., Poggi, I., van ElsWijk, G. and Wittenburg, P.: Survey of NIMM Data Resources, Current and Future User Profiles, Markets and User Needs for NIMM Resources. ISLE Deliverable D8.1, 2002b.

Marriot, A., Pelachaud, C., Rist, T., Ruttkay, S., and Vilhjalmsson, H. (Eds.). Proceedings of the AAMAS Workshop on Embodied conversational agents - let's specify and evaluate them! Held in conjunction with The First International Joint Conference on Autonomous Agents and Multi-Agent Systems, Bologna, Italy, July, 2002. http://www.vhml.org/workshops/AAMAS/papers.html

Martell, C.: FORM: An Extensible, Kinematically-based Gesture Annotation Scheme. Proceedings of the International CLASS Workshop on Natural, Intelligent and Effective Interaction in Multimodal Dialogue Systems. Copenhagen, June, 2002, 101-105.

Massaro, D. W.: The Psychology and Technology of Talking Heads in Human-Machine Interaction. Proceedings of the International CLASS Workshop on Natural, Intelligent and Effective Interaction in Multimodal Dialogue Systems. Copenhagen, June, 2002, 106-119.

Milde, J.-T.: Creating Multimodal, Multilevel Annotated Corpora with TASX. Proceedings of the International CLASS Workshop on Natural, Intelligent and Effective Interaction in Multimodal Dialogue Systems. Copenhagen, June, 2002, 120-126.

Raggett, D.: Task-Based Multimodal Dialogs. Proceedings of the International CLASS Workshop on Natural, Intelligent and Effective Interaction in Multimodal Dialogue Systems. Copenhagen, June, 2002, 127-136.

Reithinger, N., Lauer, C. and Romary, L.: MIAMM - Multidimensional Information Access using Multiple Modalities. Proceedings of the International CLASS Workshop on Natural, Intelligent and Effective Interaction in Multimodal Dialogue Systems. Copenhagen, June, 2002, 137-140.

Sidner, C. L.: Engagement between Humans and Robots for Hosting Activities. Proceedings of the International CLASS Workshop on Natural, Intelligent and Effective Interaction in Multimodal Dialogue Systems. Copenhagen, June, 2002, 141-151.

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.

van Kuppevelt, J., Dybkjær, L. and Bernsen, N. O. (Eds.): Natural, Intelligent and Effective Interaction in Multimodal Dialogue Systems. Kluwer Academic Publishers, 2003 (to appear).

Walker, M., Litman, D., Kamm, C. and Abella, A.: PARADISE: A Framework for Evaluating Spoken Dialogue Agents. Proceedings of the 35th Annual Meeting of the Association of Computational Linguistics, ACL'97, 1997.

11.2 Websites

AAMAS 2002 Conference Workshop: http://www.vhml.org/workshops/AAMAS/papers.html

Anvil coding tool: www.dfki.de/~kipp/anvil

Atlas coding tool: http://www.nist.gov/speech/atlas/index.html

CORBA: http://www.corba.org/

DARPA Communicator: http://fofoca.mitre.org

DISC project: www.disc2.dk

ELSNET roadmaps: www.elsnet.org/roadmap.html

HF 2002 Conference Workshop: , http://www.vhml.org/workshops/HF2002/papers.shtml

INSPIRE project: http://www.inspire-project.org/

ISLE workshop: http://www.research.att.com/~walker/isle-dtag-wrk/

International Conference on User Modeling: http://www2.sis.pitt.edu/~um2003/

MIAMM project: http://www.loria.fr/projets/MIAMM

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

NITE coding tool: nite.nis.sdu.dk

OAA, Open Agent Architecture: http://www.ai.sri.com/~oaa/

PRICAI 2002 Conference Workshop: http://www.miv.t.u-tokyo.ac.jp/~helmut/pricai02-agents-ws.html

SMADA project: http://smada.research.kpn.com/MainPage/

VICO project: http://www.vico-project.org

W3C, World Wide Web Consortium: http://www.w3.org/

W3C 2002 Multimodal Interaction Activity: http://www.w3.org/2002/mmi/