Natural Human-Human-System Interaction

Niels Ole Bernsen

Natural Interactive Systems Lab.

Odense University, Denmark

Abstract

The importance of a vision can be that of providing a model within which we think and create. If the model is outdated, thinking becomes unduly constrained. The paper proposes to replace the paradigm of human-computer interaction (HCI) with a more comprehensive model for thinking about future systems and interfaces. Recent progress in speech technologies has managed to establish a powerful application paradigm, i.e. that of natural task-oriented spoken language dialogue systems. This application paradigm points towards the broader goal of natural human-human-system interaction (HHSI) in virtual, combined virtual and physical, and physical environments. On the backdrop of the natural HHSI model and the rapidly changing environment of advanced systems research, the types of research that are likely to be needed in the future are discussed. The discussion deliberately de-emphasises next-generation systems research in order to shift the focus to a range of equally important, existing or emerging research objectives which sometimes show a tendency to be overshadowed by the next-generation challenges.

1. From Single Word to Spoken Dialogue

During the past 40 years or so, the field of speech technology has moved its focus from research on single word recognition to research on natural spoken human-system dialogue. The underlying tale of gradual progress in research is not the whole story, however. In those four decades, the environment in which research progress was made has changed dramatically, leading to entirely new perspectives for speech technology research. Taking a closer look at these developments may be helpful in trying to understand the roles and objectives of research in today's Information Society as well as where we are, or should be, going in the 21st century.

In 1960, promising speech recognition rates were reported for very small vocabulary (10 words), speaker-dependent, real-time recognition of isolated words [1]. Today, academic research in speech recognition seems about to reach the end of the road, being replaced by steady progress through competitive industrial development [2]. Medium-sized vocabulary (+5.000 words), speaker-independent, real-time recognition of continuous (or spontaneous) speech has become commercial reality, and very large vocabulary (+60.000 words) spoken dictation systems which only need a minimum of speaker-dependent training, can be purchased for less than 100 $US from companies such as IBM, Dragon Systems and Philips. Today, unlimited vocabulary, real-time speaker-independent continuous speech recognition is within reach and speech recognition technology has become a component technology which is finding its way into all sorts of interfaces to computer systems.

In itself, speech recognition is a transformation of the acoustic signal into an uninterpreted string of words which may or may not make sense to a human but does not make any sense to the

machine. This enables applications such as the 'phonetic typewriter' [1] as well as spoken command applications in which the system executes in response to a spoken word or phrase rather than in response to the push of a button in the keyboard, mouse or otherwise. Between humans, speech is much more than that, of course. Speech is the primary modality for the interactive exchange of information among people. Whilst hardly visible - even as a long-term goal - in 1960, the past 10 to 15 years have seen the emergence of a powerful form of interactive speech systems, i.e. task-oriented spoken language dialogue systems [3]. These systems not only recognise speech but understand speech, process what they have understood, and return spoken output to the user who may then choose to continue the spoken interaction with the machine in order to complete the interactive task. In their most versatile form, today's spoken language dialogue systems, or SLDSs, for short, incorporate speaker-independent, spontaneous speech recognition in close-to-real-time.

It is the task orientation which has made SLDSs possible at the present time. It is still much too early to build full-fledged conversational SLDSs which can undertake spoken interaction with humans the same way humans communicate with one another using speech-only - about virtually any topic, in free order, through free negotiation of initiative, using unrestricted vocabulary speech and unrestricted grammar, and so on. However, a range of collaborative tasks are already being solved commercially through speech-only dialogue with computer systems over the telephone or otherwise. One of the simplest possible examples is a system which asks if the user wants to receive a collect call. If the user accepts the call, the system connects the user. And if the user refuses the call, the system informs the caller that the call was rejected [4]. A more complex task for which commercial solutions already exist, is train time-table information [5]. The user phones the system to inquire, for instance, when there are trains from Zürich to Geneva on Thursday morning, and receives a (spoken) list of departures in return. As these examples show, task-oriented SLDSs constitute a potentially very powerful application paradigm for interactive speech technologies, which could be used for a virtually unlimited number of interactive user-system tasks in company switchboard services, banking, homes, cars etc. However, successful SLDSs remain difficult to build for reasons which go beyond the purely theoretical and technical issues involved and which illustrate the general state of speech technology research at this point.

Even if capable of working as stand-alone systems, speech recognisers are increasingly becoming components of larger and more complex systems. Likewise, speech generators which also can work as stand-alone technologies, i.e. as text-to-speech systems, are increasingly becoming system components as well. The SLDS is probably the most important technology which integrates speech-to-text, text-to-speech and other components, such as natural language understanding and generation, and dialogue management, but it is not the only one. Other integrated systems technologies incorporating some form of speech processing include speech translation systems [6], and multi-modal systems having speech as one of their input/output modalities (see, e.g., Chapter 9.3 in [3]). All of these integrated technologies represent a level of complexity which is comparatively new to the field of speech technology research. Together with the rapid increase in commercial exploitation of speech technology in general, those technologies have introduced an urgent need for system integration skills, human factors skills and general software engineering skills to be added to the skills of the groups which used to work in basic component technologies. The speech technology field, in other words, is now faced with the need to specialise software engineering best practice to speech technologies, and to do so swiftly and efficiently. This process or re-orientation has only just begun. This is why, for instance, the development of task-oriented SLDSs remains fraught with home-grown solutions, lack of best

practice methodologies and tools, ignorance about systems evaluation, lack of development platforms and standards, and so on. Only by solving problems such as these will it be possible to efficiently design and build task-oriented SLDSs which achieve their ultimate purpose: to conduct smooth and effortless natural dialogue with their users during interactive task resolution.

To summarise, today's speech technology research has several general characteristics, including:

- from components research to integrated systems research;
- one high-potential systems application paradigm (the SLDS);
- universality, speech components can be included in all sorts of systems and interfaces;
- industry and research are increasingly working on "the same things";
- researchers are facing the entire spectrum of issues of field-specific software engineering best practice, including life-cycle best practice, evaluation best practice, human factors, the need for dedicated development support tools, quality control, standardisation etc.

It is difficult to exaggerate the present significance of the above characteristics for the field of speech technology research. Together, those characteristics subsume most of the challenges facing researchers in the field. Comparison may be useful with a related technology field which also holds a strong potential for the future, i.e. that of natural language processing. Whilst the 'basic unit of research' in the speech field is the spoken dialogue, the basic unit of research in the natural language processing field is the written text. In several ways, current research is at comparable stages in speech processing and natural language processing. For instance, just as the speech field has developed mature speech recognition, the natural language processing field has developed mature spelling checkers and parsers. At the moment, however, perhaps the main difference between the two fields is that the natural language processing field has not yet developed any system application paradigm corresponding to SLDSs. There are reasons for that, of course, the principal (and deepest) one being that there is no such thing as a task-oriented text, not to speak of a task-oriented interactive text. It continues to be difficult to identify a first powerful systems application paradigm for text processing, which does not require solution to the massive research problem of processing texts in general. If such a paradigm could be found, we could expect rapid progress to be made in extremely important areas, such as task/domain-specific text translation, or task/domain-specific text summarisation. The many failed attempts at doing these things suggest that those attempts may have been incompatible with the very nature of texts which is to be unrestricted in principle, or to be unpredictably restricted which amounts to the same thing. This is not to say, of course, that highly useful text translation or text summarisation will not happen until we have mastered the processing of unrestricted text, only that one or more powerful application paradigms are still missing. The lack of a systems application paradigm for text processing means that the natural language processing field is a considerable distance behind the speech processing field with respect to having to address integrated systems research, field-specific software (systems) engineering best practice, human factors and so on.

2. From Spoken Dialogue to Natural Human-Human-System Interaction

The challenges to current speech technology research described in the previous section can be viewed in a different perspective as well. In this latter perspective, those challenges serve to generate a much more long-term vision the gradual realisation of which will transform speech technology research even more, eventually absorbing it into the wider field of natural interactive

systems. Moreover, this vision appears to be a necessary one in the rather precise sense that the vision seems to be the only possible projection from the present state-of-the-art.

Task-oriented SLDSs are not task-independent conversational systems, of course, and it is straightforward from the existence of task-oriented SLDSs to project the ulterior goal of building unrestricted conversational systems. So the creation of unrestricted conversational systems is part of the long-term vision presented here. From the point of view of technological feasibility, however, this goal is of the same magnitude and complexity as the production of unlimited text processing (understanding, translation etc.) systems. In what follows, I shall focus on some other limitations that are inherent to task-oriented SLDSs as described above, limitations which might be surmounted without having to go all the way to unlimited conversation machines.

One such limitation is the tacitly assumed human-computer interaction (HCI) paradigm: typically, a person phones the computer and conducts a spoken dialogue with it until the task has been completed (or abandoned). There is no reason why our thinking should be limited in this way. In human spoken communication, two-person-only dialogue has no privileged position or overall advantage over dialogue among three or more people. Moreover, the 'computer' part of the phrase 'human-computer interaction' has become highly misleading in the present world of networks, client-server architectures, call centres etc. A more adequate interactive paradigm for the future, then, is the human-human-system interaction (HHSI) paradigm in which two humans communicate with each other as well as with the system. An example of this setup is the Magic Lounge project which belongs to the focused research programme on computing for communities under the European Intelligent Information Interfaces (i3) initiative [7]. In the Magic Lounge, several humans meet to conduct virtual meetings among themselves with the system as note-taker, information seeker and meeting moderator [8,9]. It should be added here that, clearly, human-human-system interaction is not limited to communication in virtual space but applies to physical (or local) communication and to mixed-reality communication as well. The one system/two people configuration is the basic model. One system/one person and one system/+2 people configurations are viewed as extensions of the basic model. And the HHSI paradigm is perfectly compatible with task-oriented communication rather than unrestricted communication.

A third important limitation to task-oriented SLDS applications concerns the speech-only aspect. Humans tend to use speech-only when they communicate using, e.g., ordinary telephones, mobile phones or MBone speech-only. When humans physically meet, however, their communication tends to become far richer. When interacting face-to-face through speech, humans communicate in many other ways in parallel, using a rich set of partly redundant, partly complementary modalities for the exchange of information, including lip movement, facial expression, gesture and bodily posture, and they frequently make use of, or create, objects which are present in the environment and which themselves may have communicative contents, such as texts, maps, images, data graphics, physical models etc. As humans, in other words, we are already implementations of the *natural interaction paradigm* within which spoken dialogue is merely an, admittedly central, input/output modality among others. Researchers have begun to endow computer systems and interfaces with capabilities for natural interaction. Output lip movement synchronised with output speech has been achieved. Input lip movement recognition, output facial expression generation, and input facial expression comprehension by machine are topics for ongoing research [10]. Prototype systems exist which are capable of understanding task-oriented combinations of input speech and pointing gesture [11,12]. Gesture-only recognition by machines using cameras or other sensor systems has become a popular research topic with gesture-based music generation as an early application.

Combining the two paradigms that were projected from current task-oriented SLDSs above, we get the vision, or model, of *natural human-human-system interaction.* The task-orientation that is required today probably will go away in due course but this is not the main point. The main point is a new interaction paradigm which is capable of tremendous progress through generation of an unlimited number of increasingly sophisticated applications. In the natural HHSI paradigm, the system's role is twofold. First, the system increasingly communicates with humans the same way in which humans communicate with each other. In virtual co-presence situations, for instance, humans will communicate primarily using speech whether or not the communication is augmented with video, application sharing etc. The present chat technologies seem unnatural and are likely to largely disappear. Secondly, the system will become an increasingly all-knowing tool capable of quickly retrieving any information needed by the humans in the course of their interaction. Bottlenecks today include, i.a., bandwidth limitations on network access, software platform incompatibilities and the primitiveness of current agent technologies.

In brief, traditional human interaction for highly generic purposes, such as problem-solving, is characterised by:

- X people physically together, documents, drawings, physical models, phone, fax, oral discussion, gesture, facial expression, emotion, sketching, demonstrating how to do, etc.
- Natural human-human communication, in one place, with limited access to external knowledge.

Human problem-solving in the future will be characterised by:

- X people together in physical and/or virtual space.
- The system is an (increasingly) all-knowing tool.
- The system is a natural communication partner.
- Natural human-human-machine communication, ubiquitous, unlimited knowledge access - for all users.

Obviously, joint problem-solving is not the only broad family of tasks which will be ”taken over” by the natural HHSI paradigm. Game-playing, for instance, will be so as well. The phrase ”for all users” is important. Enabling natural communication with machines will serve to reduce to its proper role the GUI (graphical user interface) paradigm in which interaction is being done through mouse-like input devices, keyboard input and screen output. Instead, ubuquitous systems use will increase, the classical computer will increasingly fade from our environment, and computing will no more assume the literacy and dexterity that tends to be required by GUI interfaces.

3. Future Research Challenges

The natural HHSI paradigm is easy to comprehend, seems to follow-by-projection from current system and paradigm limitations and current development trends in an almost deterministic manner, and has tremendous potential. However, like the purely technical development from single word speech recognisers to task-oriented spoken dialogue systems described at the start of Section 1, this way of presenting natural HHSI fails to consider the changing context of research, the surprises it may hold for traditional points of view, and the challenges it poses to the selection of research directions to undertake or fund. This wider context and the challenges it offers are discussed in the present section.

3.1 Mainstream Technology Research

It seems that research is moving away from focusing on basic components. There is still basic components research to do, of course, but even what remains of basic components research is increasingly being influenced by demands stemming from systems research and general software engineering requirements. Increasingly, the natural HHSI paradigm demands research into high-complexity systems, integration of several components, APIs development etc. Researchers who want to compete in the race to invent and build the next system generation(s), in other words, will need to work in the context of large and complex systems. The problem posed by this development is not so much that many researchers are not used to doing that. Rather, it means that it is no longer possible to ”build what one needs from scratch”. To do so simply is infeasible in terms of the resources needed as well as, in most cases, extremely inefficient or even silly. Instead, researchers, just like industry, have to select the components they do not want, or need, to build, from off the shelf. And the shelf itself is growing larger by the minute. Knowing what's on the big shelf is quickly becoming very difficult, the growing risk being that the research team might spend years of effort re-inventing the wheel. Furthermore, selecting from the big shelf is not always for free, which means rapidly increasing demands on financial resources.

Moreover, to be sensible, systems design and specification based on what is on the big shelf needs to take into account emerging platforms, standards and even market trends. If any one of these factors evolve differently from what was expected when the research system prototype work was launched, large amounts of effort may have been wasted. Similarly, if an expected, emerging platform or program version gets delayed, the research system prototype work may fail to reach its objectives.

These points suggest that research system prototype work following the natural HHSI paradigm involves high-risk, opportunistic technology research. That the research is high-risk and potentially resource intensive has been made clear above. That the research is opportunistic is partly coincidental with the high risk of advanced systems development in research, partly due to the fact that failed expectations concerning how the underlying technologies will develop will not necessarily lead to research project failure. Sometimes, it will be possible to re-direct efforts to follow leads that were discovered underway, thereby avoiding total failure. It is difficult to guess how often this will be the case, but clear that drastic project re-orientation imposes strong demands on the inventiveness and flexibility of the research team(s) involved. It is much easier to continue to follow an effectively dead plan than to re-orient in a flexible and opportunistic manner.

The natural HHSI paradigm, in other words, poses considerable challenges to systems research. The conditions under which this research is being carried out strongly resemble those faced by industry. The difference, however, is that innovative systems research takes place without most of the safe-guarding infrastructure of sound industrial R&D laboratories, including specialisation to one or a few range of products, substantial in-house platform resources created through past efforts, professional awareness of standards and market developments, etc. It is hard to see how these challenges to research are likely to be met in the future, unless one assumes either (a) few, large, non-industrial advanced research laboratories, (b) much more effective collaborative research than we are used to from the past, with real synergy among highly specialised teams covering all of the crucial aspects of the system technology to be built, or (c) that industry more or less takes over the entire research and development process, leaving non-industrial research

teams to do something else. Some may continue the work that remains to be done in components research. Others may do work of the kinds described in Sections 3.2 - 3.6 below.

In Europe, there is no significant tradition for (a) above. (b), however, is currently being piloted in i3, the Intelligent Information Interfaces initiative [7]. i3 features groups (or research programmes) of +10 research projects in a focused research area. The research projects start simultaneously, run for 2-3 years in parallel and have strong incentives for cross-project collaboration. The two ongoing i3 research programmes address computing for local and virtual communities, and computing in schools for the 4 to 8 years old, respectively. i3 will be followed, in the fall of 1999, by a third, related research programme called Universal Information Ecosystems [13]. As for (c), there are indications that some of the larger European IT/Telecoms are turning sceptical about the development of advanced systems prototypes in non-industrial research laboratories.

3.2 Futuristic Scenarios of Use

In some advanced systems and interfaces research areas, the picture conveyed in Section 3.1 is already commonplace. This may be particularly true of core software (and hardware) research areas. However, it may be suspected that the picture may be less familiar to many of the research teams who are best poised for adopting the natural HHSI vision. If these groups are inclined to hesitate in joining the race for the next system generation(s), or at least inclined to diversify their research pursuits, the wider context of the natural HHSI paradigm would seem to offer plenty of novel opportunities. Many of these opportunities may appear to a traditional point of view to represent a stretching of the classical concept of research - but so much the worse for the classical concept! In fact, some of us always felt slightly uneasy about academic researchers who fell victim to the ”see my beautiful system” condition. When asked for underlying theory or theoretical implications, those researchers had nothing to say except that it was interesting that their system could do what it did, wasn’t it? Arguably, researchers should not, as a group, merely build advanced systems but also generalise what they discovered while doing so.

One way of facing the tough demands on development of next-generation complex systems following the natural HHSI paradigm, is to work several generations ahead, aiming at concept demonstrators rather than complete working systems. This raises other difficulties, to be sure, but at least it frees the researcher from having to face most of the hard issues described in Section 3.1. Rather, the starting-point becomes that of designing the future lives of people. User needs and social trends come into focus, replacing the question of what might be an example of the next system generation given the state-of-the-art in products, prototypes and standards. Viewed from a high level of abstraction, user needs do not change at all. People’s needs for information, transportation, shelter, entertainment etc. remain constant throughout history. What changes are the ways in which new technologies could satisfy those needs in the context of the enormous complexity of technological and societal developments. Futuristic use scenario research seems likely to strongly increase in importance in the coming years. And, once a future scenario of use has been identified, questions arise as to how people will behave, what they will prefer and why they will do that. i3 research, which is probably some of the most long-term research done in Europe at the moment, clearly illustrates this trend towards invention of the future based on future scenarios of use and investigation of the lives of ordinary users of all kinds, from little kids to the elderly, and from all cultures.

In futuristic use scenario-based research, the technology demonstrators to be developed are likely to be, in terms of person/years of technology development, relatively modest concept

demonstrators which show how something could be done in the future through innovative technical solutions, without worrying about available platforms, existing or emerging standards etc.

However, the demonstrators might also be something else entirely, namely innovative, carefully designed product illustrations which are uniquely based on today's technology. In other words, many companies could build and market those things right away from off-the-shelf components. The point of this research is not to innovate the technology per se but to innovate the design, the technology's role in people's lives and/or the intended user population. Is this "research"? I don't know. Yet innovative product designs are certainly an important form of innovation to be expected from futuristic use scenario-based research, especially from the kind which integrates technology, design and people as in i3.

3.3 Specialised Best Practice

The need for specialised software engineering best practice methodologies for SLDSs that was noted in Section 1, can be straightforwardly generalised to systems for natural HHSI. Work on best practice for SLDSs is ongoing at the moment [14,15]. For natural HHSI systems more generally, nothing exists beyond general software engineering best practice, ISO standards and the like. There are no specialised software engineering best practice methodologies at all. This means that substantial research efforts are needed.

3.4 Efficient Data Handling

The task-orientation of much natural HHSI research implies the need for huge amounts of data. The task-orientation of such systems is due to the fact that general natural interaction systems are not likely to be built in the foreseeable future (cf. Section 1). Task-orientation implies that systems must be carefully crafted to fit human behaviour in the task domain in order to work at all. This requires deep understanding of human behaviour in the task domain. If one looks to speech recogniser and spoken language dialogue systems (SLDSs) development projects, the amounts of data needed to make these technologies succeed have been staggering. And when research is now beginning to address speech and gesture combinations for task resolution, or speech and gesture and facial expression combinations for task resolution, data capture will necessarily continue to be a major activity. This data deals with how humans actually behave when communicating in those ways and there is presently no shortcuts available for dispensing with data from experimentation, simulation, user testing, and from the field when the corresponding systems are to be developed.

Data capture, of course, is only a first step. Upon capture, the data needs to be marked up electronically, analysed, used for development and, whenever possible, re-used. In the speech recognition field, there are now standards for the data which are needed, i.e. for the amount, structure and quality of the data needed for the automatic training of speech recognisers for new languages. However, in the far more complex field of SLDSs, data standards do not yet exist. Moreover, for many increasingly important types of data, such as data on the speech (or dialogue) acts which people execute when interactively performing a task with the system through spoken language dialogue, the underlying theories and theoretically motivated concepts needed for identifying the appropriate phenomena of interest in the data, are not yet in place. This implies a need for theory which will be discussed in Section 3.6.

Finally, to handle data efficiently, software tools are needed to electronically mark up, query, visualise, import and export the data. In the field of markup (or data annotation) tools, global standards barely exist at the moment. This means that each research team or group of industrial developers marks up their own data in inefficient ways, often lacking appropriate tools, and using idiosyncratic formalisms for their purposes. This situation is only now being addressed in the case of spoken language dialogue data, for instance in the MATE project on Multilingual Annotation Tools Engineering [16]. When it comes to natural HHSI applications more generally, such as those requiring speech-and-video data conceptualisation, annotation and analysis, for instance of the communicative gestures accompanying speech, there is even far more virgin territory to be explored and cultivated.

There are several rather obvious reasons why progress has been severely lacking in the general field of the handling of data on human communication behaviour. One is that the need for development efficiency is relatively recent, the first not-quite-simple SLDSs having been developed only recently. Another is that industry is not necessarily strongly motivated to develop tools and standardisation for re-use in a field in which data used to be proprietary. A third reason is that the field used to be considered one of rather esoteric research. Today, the needs for re-usable data, efficient data handling tools and global standards have finally become clear. This implies a need for very substantial research if the natural HHSI paradigm is to be realised without undue difficulty. In view of the number of languages and cultures to be mastered, an analogy with the current high-profile genome projects, human and otherwise, comes to mind.

3.5 Design Support Tools for Usability

The field addressed in this section, i.e. that of design support tools and, in particular, design support tools for usability, still remains more of a dream than tangible reality in what used to be called human-computer interaction (HCI) research. Current natural HHSI systems must be carefully crafted to fit human behaviour in order to work at all. The capture, markup and analysis of data on human behaviour tends to be very costly, for several reasons. Some of these were noted in Section 3.4, i.e. the lack of theory, concepts, standards and markup tools. Another reason is equally important. Truly realistic data on user behaviour can only be produced from field trials with the implemented system. The Wizard of Oz simulation method [3] is useful for data capture during early design and prior to system development, but this method is far from sufficient to ensure the generation of fully realistic data. And if the field trials demonstrate serious system flaws it may be necessary to start all over again. In other words, it would seem highly desirable to have design support tools for usability which, during early design and before implementation has begun, could ensure that the system to be built will not turn out to be fatally flawed when tested in the field towards the end of the project. In the early 1990s, I participated in the Esprit long-term research project AMODEUS [17]. AMODEUS was probably the largest-scale basic research project there ever was in what used to be called HCI. To me, at least, the main outcome of AMODEUS was that design support tools for usability are (a) difficult to do (AMODEUS never developed a single such tool for actual use by developers), and (b) probably the best that HCI or, rather, HHSI research could do for system developers. It appears correct to say that few tools of this kind have been developed in the 1990s.

Stubbornly adhering to the main outcome of AMODEUS, I have continued with colleagues to explore opportunities for developing design support tools for usability. Two such tools are now about to appear after several years of work, both having been built in the DISC project on spoken language dialogue systems best practice in development and evaluation [15].

One tool supports the development of cooperative spoken system dialogue for SLDSs [18]. The basic idea is a simple one. Task-oriented spoken dialogue is undertaken to complete a particular task with a minimum of hassle. To do that, the interlocutors (human(s) and the system) should conduct shared-goal dialogue, the shared goal being that of completing the task. Such dialogue demands full dialogue cooperativity of the interlocutors. The user is not the problem. For one thing, human users implicitly know how to be cooperative in dialogue. For another, if they refuse to cooperate, they will not get their interactive task done and there is nothing we, as system developers, can or should do about that. However, if the system's dialogue is consistently cooperative, following a more or less complete set of principles of cooperativity, there is every chance that the dialogue will run as smoothly as possible, avoiding the need for clarification and repair sub-dialogues which are still rather difficult to handle by machine. The cooperativity tool, then, supports the identification, during early design, of flaws in the design of the system's dialogue contributions.

The second tool supports modality choice in the early design of complex systems which include speech as one of their modalities. As systems and interfaces depart from the GUI paradigm, developers are faced with a growing diversity of implementable ways of exchanging information between systems and their users. A simple example is that of using output speech and output data graphics together. However, much too little is known about the, always limited, functionality of available input and output modalities. Even in the apparently simple case of speech-only, the developer runs an important risk of using speech for input to, and/or output from, a system for which speech is not appropriate at all, or is not appropriate in the form in which it is planned to be used. Speech functionality has turned out to be a very complex problem. However, based on Modality Theory [19] and analysis of large sets of claims about speech functionality derived from the literature, it has turned out that a small set of basic properties of speech is sufficient to provide guidance for developers in the large majority of cases they are likely to face. The speech functionality tool provides this guidance during early design [20]. Obviously, the proper understanding of speech functionality is but part of the much larger problem of understanding all possible modalities for the exchange of information between humans and machines. This is a problem for future research.

Both of the above tools are meant to be used by system developers during early design, and might help avoid early design decisions that might later prove fatal to the usability of the implemented system. It may be worth mentioning some reasons why design support tools for usability have not been developed to any greater extent so far. One reason is that much HCI research still tends to take place too far from actual systems development. It is very hard if not impossible to develop useful tools for systems developers if one is not deeply familiar with systems development oneself. Research in the natural HHSI paradigm, it is proposed, should not commit the same mistake of divorcing the study of usability from actual systems development practice. A second reason is that design support tools for usability need theory, and HCI has not been particularly effective in developing theory. Research within the natural HHSI paradigm, it is proposed, should do better than that (see below). A third reason is that, even with a useful theory in hand, developing design support tools for usability tends to be quite time-consuming to do, primarily because of the iterative tools testing involved but also because it demands the capability for thinking in terms of educational systems design. The latter is a particular skill which is not necessarily present in someone with a useful theory for backing up an early design support tool.

3.6 Useful Theory

The scarcity of applicable theory pertaining to the natural HHSI paradigm has been noted above. However, before addressing the topic of missing and strongly needed natural HHSI theory, it may be appropriate to inquire about the status of theory in the world of advanced IT/Telecom research more generally. I recently asked a project officer working under the European Commision's huge 5th Framework Programme's 4 Billion $US Action Line on Information Society Technologies, if the term 'theory' was mentioned anywhere as something that might be funded in research projects supported by the programme. He answered that he didn't think so. Ten years ago, the corresponding 2nd Framework Programme actually did fund theoretical work as part of its long-term research branch. For instance, funding was provided for core computer science topics, such as Complexity Theory and Petri Nets, and for HCI research. Meanwhile, core computer science has become somewhat marginalised and general HCI research has been dropped largely because it failed to deliver to a satisfactory extent. Obviously, however, facts such as these do not imply that highly relevant theory is not needed, or feasible, any more.

I would like the following discussion to serve as a plea for basic theory. It is possible, of course, that nobody disagrees with the argument below and that basic theory was just forgotten in FP5. Or it may be that everybody agrees but does not view basic theory development as part of the long-term research to be funded by the European Commision's research programmes. There are reasons for endorsing the latter view, to be sure. It can be easy, and hence tempting, to "over-sell" theory by stressing its application potential but forgetting the time it takes to develop the theory itself as well as the time it takes to render it applicable. Furthermore, theory is often done by single individuals, which makes its development less amenable to funding through collaborative research programmes. Whatever the prevailing thinking, here follows a plea for practically useful theory.

The 1980s with their visions of unifying research programmes in artificial intelligence, cognitive science, HCI etc., are long gone. In the practical and entrepreneurial 1990s, general software engineering has tended to form the only centre of our work, the rest of which has been about creating innovative technical solutions. As argued above, the scene is rapidly changing once again, towards the union of technology, design and people, towards creative contents, cultural diversity and so on. From the point of view of practically useful theory, the focus on providing next-step technical solutions ignores the progress that has been made towards increasingly sophisticated interactive technologies.

Let us, once again, take speech technology as an example. Task-oriented spoken language dialogue systems (SLDSs) research is now reaching into hitherto rather obscure research areas traditionally belonging to the arts and humanities, such as speech act theory, co-reference resolution theory, cooperativity theory, politeness theory, the theory of cultural differences in the way information is being exchanged etc. To quote just one example, the huge eight-years, 160 Mio. Deutchmarks German national project in task-oriented spoken translation by machine, Verbmobil [6], could not have proceeded without making machines able to identify and process the speech acts involved in the task chosen for the project (appointment scheduling between humans). In response, the Verbmobil researchers had to create a speech acts taxonomy from scratch, and this taxonomy remains one of the few major efforts in applicable speech acts analysis worldwide [21]. The interesting question is: why could the Verbmobil researchers not just fetch the theory they needed from the big shelf in the arts and humanities field? The answer is that, so far, the arts and humanities have not at all been geared to providing technologically

applicable theory. The Verbmobil researchers could fetch the generic concept of speech acts (or dialogue acts) from their shelf but that was all.

Let us add to the above line of argument a more general observation. In developing tomorrow's SLDSs, we are actually facing the task of 'reconstructing', from the bottom up, the entire field of human spoken communication. As the state-of-the-art in arts and humanities research cannot deliver except sporadically, we have to start doing this ourselves, thereby creating what potentially is a tremendously productive interface to entirely new disciplines as viewed from the world of IT/Telecoms research. The work on cooperativity theory mentioned in Section 3.5 is a case in point. We found that by analysing real data from human interaction with a spoken language dialogue system, it became possible to significantly augment an existing 'theoretical island' in arts and humanities research, i.e. that of Grice's cooperativity theory [22,23]. The results of this work have now been incorporated into an early design support tool for SLDSs developers. The general point just illustrated is that research into non-technological theory is strictly and provably needed for technological progress. Another example comes from the development of annotation tools for the markup of spoken language dialogue data. To mark up some corpus of data, one needs concepts of the phenomena to be marked up. These concepts, such as 'subject' and 'object' in syntactical annotation or the concepts of different types of speech act, comes from the corresponding theory which has the task of providing some form of closure and rationale why the conceptualised phenomena are those and only those to expect in a certain segment of natural spoken dialogue. Theory is needed for corpus markup for exactly the same reason why enabling the machine to recognise those phenomena is needed.

Now, generalising in earnest, so to speak, beyond current research issues, the advent of natural HHSI systems including the full combinatorics of speech, lip movements, gesture, facial expressions, bodily posture, intentions communication, emotions communication, and inherent reference to all sorts of environmental objects - is not only likely, but certain to produce an increased need for theory about human behaviour. Note, incidentally, that this is not about psychology, about what goes on in people's minds, but about objectively observable entities and regularities in data from natural human-human-system interaction. It is strictly necessary to be able to conceptualise those entities and regularities in order for machines to identify and use them, and in order to be able to mark them up in the data analysis in order to train the machines to identify them.

What the above argument suggests is that the building of tomorrow's natural HHSI machines requires what amounts to a complete re-building of the theory of human-human communication from the bottom up with the system thrown into the loop. Without it, we shall not be making the progress which we can very easily envision through simple intuition, as demonstrated by the intuitive convincingness of the natural HHSI vision.

Shifting the topic to the virtual and/or physical co-presence of humans and machines, the same picture emerges. Social theory has technologically relevant 'islands of theory' concerning human co-presence in meetings and otherwise. We are now actually building virtual co-presence systems. To do that, we need solid knowledge about the needs and behaviours of humans in group encounters aiming at the resolution of particular tasks. Existing social theory is far from being able to deliver everything we need because it has been developed with different, less technologically focused objectives in mind. However, to plausibly specify a virtual meeting support system we need all the theoretical knowledge we can get. If we cannot find it on the shelf in the social sciences, we must begin developing that knowledge ourselves. The needs of

technology are likely to enforce a reconstruction, from the bottom up, of social theory. Otherwise, we will not be able to make sense of the data we collect on human-human-machine interaction and, maybe even more striking, we cannot make the systems which we painstakingly build, usable by their intended users.

In a final example, Modality Theory, which was used as a foundation for building the speech functionality tool mentioned in Section 3.5, was not imported from the arts and humanities but was developed from scratch in the context of anticipating the needs for applicable theory in the emerging world of multimodal systems and interfaces [24].

3.7 Implications for Future Research

This section looks at implications for some of the key issues to be discussed at the workshop. Natural HHSI is not a magic solution to the quest for perfect interaction with all systems and for all users. Pending progress in other fields, users may still get lost in virtual space, miss helpful context-sensitive help and intelligent adaptation to their user profile, or find multimedia systems, however advanced their graphics, boring, stereotypical and lacking in opportunities for creative interaction. The natural HHSI paradigm as presented in this paper remains limited by the task-orientation and focused primarily on the - admittedly very broad - area of joint human-human-system problem-solving. If one transcends those boundaries, for instance by assuming natural conversational capabilities of systems displayed as fully natural avatars, many novel opportunities emerge, such as cracking jokes together or discussing the war in Kosowo. If and when that becomes possible, however, other fields of research will be likely to have produced truly entertaining interactive systems as well.

How to reduce cognitive load and provide more scope for creativity? The creativity issue was briefly addressed in the preceding paragraph. Natural communication is by nature economical with respect to cognitive load. In general, this is because it is replete with information redundancy conveyed through speech, lip movements, facial expression, gesture and bodily posture. The human system reads off this information in parallel and is able to patch up missing information in one modality by using redundant information from other modalities. As for system input, the system's capability of understanding spontaneous communication removes any need for remembering particular spoken or written keywords, lip movements, facial expressions, gestures, actions or bodily postures.

Cross-disciplinary interaction and how to make it work. After nearly two decades of rhetoric concerning the need for interdisciplinarity in advanced systems development, it may be timely to point out that, at least in the general field of natural interactive systems, interdisciplinarity is actually working. In fact, it is working so well that the rhetoric of the past has begun to appear misleading. Today, much research could be described rather as post-disciplinary, being conducted by individuals whose combined training and professional experience makes obsolete the question which particular discipline they represent. When such people work together, their collaboration is no longer interdisciplinary in any recognisable sense.

Ways of handling interaction in specific social contexts and differences of culture. Addressing social and cultural differences is inherent to the natural HHSI paradigm because of the fact that natural communication varies widely due to those differences.

Dealing with universality and the problems of the differently-abled. The redundancy among the different modalities in natural communication is far from being complete. This poses challenges for natural interactive technologies aimed at substituting one modality for another when

communicating with differently-abled users. In most languages, for instance, not all spoken phonemes can be distinguished through lip reading. One solution is to add simple haptic output code which enables the deaf to disambiguate those phonemes. This may serve to illustrate the potential of multimodal technologies for achieving full expressiveness in communication with the differently-abled.

Interaction styles and their implications. By definition, (fully) spontaneous, natural input is the ultimate in human-to-system communication because humans cannot communicate any better than that in terms of expressiveness. In the future, we might even want to be able to selectively prevent the system from understanding all aspects of our natural communication. In terms of devices, cameras and microphones are able to capture all relevant graphical (-to-the-system) and acoustic information. As for haptic input, we will be seeing a proliferation of ergonomic input devices that are as many and as varied as those produced throughout history for non-electronic use (i.e. doorknobs, forks etc.). The system output space is an altogether different matter. Apart from being manageable through the input devices available, and apart from requiring that the system use output speech, gesture, etc. whenever appropriate, no further naturalness requirement would seem relevant to system output behaviour. This is exemplified below.

Consistency of cognition models across information appliances. This is a major challenge not only for the natural HHSI paradigm but for interaction design in general. We would like to be able to chart the cognitive implications of using any possible modality combination as input and/or as output during interaction. There is a clear need for applicable theory on this issue. It was argued in [20] that there is no way the issue can be solved through systematic empirical investigation because of (i) the sheer complexity of the issue itself and (ii) because of the failure of HCI research to deliver basic conceptual instruments, such as general task taxonomies, user taxonomies, cognitive property taxonomies, systems taxonomies etc. The approach proposed in [20], which has been corroborated in a recent control study, is that of charting the basic cognitive implications of using different modalities. Based on Modality Theory [19, 24], it is possible to identify all relevant individual modalities. Based on this individuation, it becomes possible to encapsulate the cognitive implications of using each individual modality in terms of 'modality properties', such as that "Speech input/output modalities, being temporal (serial and transient) and non-spatial, should be presented sequentially rather than in parallel" or that "Static graphic/haptic input/output modalities allow the simultaneous representation of large amounts of information for free visual/tactile inspection and subsequent interaction" [20]. Even if the challenge remains formidable, the hypothesis confirmed through those studies is that we may need only, perhaps, 100-200 modality properties to tackle a complexity of interaction which may be estimated to lie in the range of billions of combinations of the relevant parameters. A less encouraging result was that a particular class of important HHSI issues was not likely to become theoretically transparent through the described approach. These are issues to do with input speed, such as whether speech input is slower than push-button input for time-critical interaction in aviation, or whether combined speech and pen interaction is faster than both pen-only interaction and speech-only interaction for generic tasks, such as digital roadmap interrogation or text editing. In general, the more complex and detailed (in terms of the number of parameters and the subtlety of distinctions involved) an issue of modality choice becomes, the less likely it is that the issue can be resolved theoretically at all. However, if theory can help resolve those +90% issues which are less complex and less subtle than that, this is still important progress. For the rest, and depending on the business plan economy and the willingness to take a high risk, developers will have to initiate scientific experimentation.

What are the paradigms for emerging new kinds of interaction - beyond WIMP interfaces: multimodal and perceptual user interfaces. The natural HHSI paradigm.

Current challenges for virtual environment technology and interfaces. Coupling task-oriented natural interaction with virtual environments may seem to generate a paradox. In virtual environment graphics games, for instance, displayed natural actor/environment behaviour is often at a premium. However, when solving particular practical tasks in virtual environments, efficiency will often be at a premium. However, efficiency is not always compatible with displayed natural actor/environment behaviour. When shopping in a virtual department store, for instance, few users would want to watch their mouse pointer, or their avatar, majestically ascending the escalator to the second floor. In other words, we do not necessarily want virtual reality to behave naturally when a task has to be done. This poses important challenges for combining virtual environments and task-oriented user behaviour.

Usability issues and measuring the effectiveness of symbiosis. As argued in connection with the issue of "consistency of cognition models across information appliances" above, it is important to have to rely on empirical experimentation only in cases of relatively extreme complexity of the parameters involved. Modality Theory acknowledges about 60 unimodal modalities. We currently acknowledge 15 'domain variables', such as 'generic system', 'generic task', 'performance parameter', 'user group', 'learning parameter' or 'cognitive property', each of which has on average, say, 25 instantiations, such as 'cognitive property [limited retainability]'. The total complexity can be much higher than that - just add, for instance, a combination of several modalities - but this complexity is already far beyond systematic resolution through scientific experimentation in which each modality and (each domain variable instantiation minus one, the dependent variable) has to be fixed in advance, and the results of which can be safely generalised only to a very limited extent. In fact, the experimental paradigm for HHSI (or HCI) already died around 1991 as can be seen from the dramatic drop in accepted experimental HCI papers for the CHI Conference in those years. In the complex field of multimodal input/output, in other words, 'measuring' should be limited to (a) very complex modality issues for which any theory is bound to have close-to-zero predictive power and (b) standard software engineering evaluation practice performed as part of the development life-cycle. What theory can do is to limit the need for measurement with respect to most modality issues (cf. a) and limit the risk of developing unusable artefacts (cf. b).

4. Conclusion

The natural HHSI vision has been proposed in this paper as a model which might be useful for thinking about future research on systems and interfaces. However, instead of discussing at length which those families of systems and interfaces might be, the paper has deliberately de-emphasised next-generation IT/Telecoms demonstrators and applications. Instead, a broad look was taken at future lines of research which appear to be needed independently of the particular nature of the natural HHSI systems which will be developed in the coming years. The result was a number of research directions including the exploration of the future lives of people, futuristic concept demonstrators, specialised software engineering best practice methodologies, data handling schemes and tools, design support tools for usability and, last but not least, a renewed emphasis on theory development. Underlying these directions of research is something else, i.e. the need for systems development to forge strong links with hitherto remote areas, such as design, arts and humanities, social theory, and prospective users of all kinds and from all

cultures. This is not just a matter of speech technologists talking to natural language processing researchers or to machine vision researchers. I don't think that the term 'interdisciplinarity' which has been around in our fields for more than fifteen years, and which has begun to carry the same antiquarian connotations as, e.g., the term 'modern', adequately captures those contemporary and future needs. What we are looking towards, is the *post-disciplinary* world of tomorrow's research for the Information Society, in which researchers work along lines such as those described above, and on the basis of knowledge and experience which is far from that represented by any known classical discipline.

References

[1] Fatehchand, R.: Machine recognition of spoken words. In F. L. Alt (Ed.): *Advances in Computers,* Vol. 1. New York: Academic Press 1990, 193-229.

[2] *Elsnews,* The Newsletter of the European Network in Language and Speech, 8, 1, 1999.

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

[4] Bossemeyer, R. W. and Schwab E. C.: Automated alternate billing services at Ameritech: speech recognition and the human interface. *Speech Technology Magazin*e 5 (3), 1991, 24-30.

[5] Aust, H.: The Philips automatic train timetable information system. *Speech Communication* 17, 249-262, 1995.

[6] Wahlster, W.: Verbmobil - Translation of face to face dialogues. *Machine Translation Summary IV,* Kobe, Japan, 1993.

[7] http://www.i3net.org

[8] http://www.dfki.de/imedia/mlounge/

[9] Bernsen, N. O., Rist, T., Martin, J. C., Hauck, C., Boullier, D., Briffault, X., Dybkjær, L., Henry, C., Masoodian, M., Néel, F., Profitlich, H. J., André, E., Schweitzer, J., Vapillon, J.: Magic Lounge: A thematic inhabited information space with "intelligent" communication services. In Rault, J.-C. (Ed.): *La Lettre de l'Intelligence Artificielle, Proceedings of the International Conference on Complex Systems, Intelligent Systems, & Interfaces (NIMES'98),* Nimes, France. Nimes 1998, 188-192.

[10] Cole, R. A., Mariani, J., Uszkoreit, H., Varile, G., Zaenen, A., Zampolli, A. and Zue, V. W. (Eds.): *Survey of the State of the Art in Human Language Technology,* Chapter 9. URL: http://www.cse.ogi.edu/CSLU/HLTsurvey/

[11] Goddeau, D., Brill, E., Glass, J., Pao, C., Phillips, M., Polifroni, J., Seneff, S. and Zue, V. W.: Galaxy: A human-language interface to on-line travel information. In *Proceedings of the ICSLP94,* Yokohama, 1994, 707-710.

[12] Guyomard, M., Le Meur, D., Poignonnec, S. and Siroux, J.: Experimental work on the dual usage of voice and touch screen for a cartographic application. In *Proceedings of the ESCA Workshop on Spoken Dialogue Systems,* Vigsø, Denmark, 1995, 153-56.

[13] http://www.cordis.lu/ist/fetuie.htm

[14] Dybkjær, L., Bernsen, N. O., Carlson, R., Chase, L., Dahlbäck, N., Failenschmid, K., Heid, U., Heisterkamp, P., Jönsson, A., Kamp, H., Karlsson, I., Kuppevelt, J. v., Lamel, L., Paroubek, P., Williams, D.: The DISC approach to spoken language systems development and evaluation.

In Rubio, A., Gallardo, N., Castro, R. and Tejada, A. (Eds.): *Proceedings of the First International Conference on Language Resources and Evaluation,* Granada, 1998. Paris: The European Language Resources Association, 1998, 185-189.

[15] http://www.elsnet.org/disc/

[16] http://mate.mip.ou.dk/

[17] http://www.mrc-cbu.cam.ac.uk/amodeus/

[18] Bernsen, N. O., Dybkjær, H. and Dybkjær, L.: What should your speech system say to its users, and how? Guidelines for the design of spoken language dialogue systems. *IEEE Computer,* 30 (12), 1997, 25-31.

[19] Bernsen, N. O.: Defining a Taxonomy of Output Modalities from an HCI Perspective. *Computer Standards and Interfaces,* Special Double Issue, 18, (6-7), 1997, 537-553.

[20] Bernsen, N. O.: Towards a tool for predicting speech functionality. *Speech Communication* 23, 1997, 181-210.

[21] Jekat, S., Klein, A., Maier, E., Maleck, I., Mast, M. and Quantz, J.: Dialogue acts in VERBMOBIL. *Verbmobil Report* 65, Universität Hamburg, DFKI Saarbrücken, Universität Erlangen, TU Berlin, 1995.

[22] Grice, P.: Logic and conversation. In P. Cole and J. L. Morgan (Eds.), *Syntax and Semantics* Vol. 3: *Speech Acts*. New York: Academic Press 1975, 41-58.

[23] Bernsen, N. O., Dybkjær, H. and Dybkjær, L.: Cooperativity in human-machine and human-human spoken dialogue. *Discourse Processes* 21, (2), 1996, 213-236.

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

Biography

Niels Ole Bernsen is director of the Natural Interactive Systems Laboratory at the University of Southern Denmark - Odense. He was trained as a philosopher, took his PhD equivalent in epistemology and his dr. phil. degree in the theory of cognitive situations. He worked in Brussels 1986-1989, first as a researcher investigating future perspectives of cognitive science and later assisting in launching the Esprit Basic Research Actions in microelectronics, computer science, artificial intelligence and cognitive science. He went back to Denmark as a State Budget research professor in cognitive science working in HCI and connectionism. Gradually, this work has changed into systems engineering and he is now a professor of engineering at Odense University working in natural interactive systems, design support tools, modality theory and systems evaluation. He is the Coordinator of the European Network for Intelligent Information Interfaces (i3net) which supports 25 long-term research projects in computing for virtual and local communities and in experimental school environments. He is a member of the Executive Board of the European Language and Speech Network (Elsnet).