# SPEECH-RELATED TECHNOLOGIES

# Where will the field go in 10 years?

**Niels Ole Bernsen, NISLab, Denmark (editor)**

## Abstract

This paper is a draft position paper for discussion at the ELSNET Brainstorming Workshop 2000-2010 in Katwijk aan Zee, The Netherlands, on 23-24 November, 2000. The paper first describes some general emerging trends which are expected to deeply affect, or even transform, the field of speech technology research in the future, including trends towards advanced systems research, natural interactivity, multimodality, and medium-scale science. A timeline survey of future speech-related technologies is then presented followed by analysis of some of the implications of the proposed timelines. Timeline projections may turn out to have been false, of course, but even their turning out to be true is subject to future actions which are (not) taken to make them true. Accordingly, the final part of the paper discusses some actions which would seem desirable from the point of view of strengthening the position of European speech-related research.

## 1. Introduction

The term *speech-related research* has been chosen to designate the topic of the present paper for lack of ability to invent a more appropriate term, if there is one. At least, the term partly manages to convey the author's expectation that the field of speech research will change rather dramatically in the coming ten years as speech technologies become merged with other technologies into a field which, so far, lacks a name.

According to many observers, the coming decade will be the decade of speech technologies. Computer systems, whether stationary or mobile, wired or wireless, will increasingly offer users the opportunity to interact with information and people through speech. This has been made possible by the arrival of relatively robust, speaker-independent, spontaneous (or continuous) spoken dialogue systems in the late 1990s as well as through the constantly falling costs of computer speed, bandwidth, storage, and component miniaturisation. The presence of a speech recogniser in most appliances combined with distributed speech processing technologies will enable users to speak their native tongue when interacting with computer systems for a very large number of purposes. Although no doubt exaggerated as just presented, there probably is some truth to this vision of a breakthrough in the application of speech technologies in the coming years. If this is the case, it would seem worthwhile that we lift our sights and take a long-term view of the issues ahead. This may help setting a reasonable research agenda for the coming years of advanced speech systems research and development, one which does not succumb to the usual hype associated with fashionable technologies. Today, some believe that "the speech problem" has been solved already. Some believe that speech, because of its naturalness, is the solution to every conceivable problem of user-system interaction. On the other hand, surprising as it may seem, some human factors and interactive systems experts believe that we have just arrived at the touch-tone telephony stage and share no notion of the actual state-of-the-art in the field with its practitioners. Since

all of those beliefs are far from the truth, it is important to provide a more balanced picture of the state-of-the-art in speech technologies in order to set the stage for solid progress.

In what follows, Section 2 presents some trends in the speech-related research field. Section 3 excels in guesswork by estimating the times of appearance of a range of novel speech-related technologies. Section 4 discusses implications of the timelines presented in Section 3. Section 5 proposes a series of actions which would appear appropriate given the preceding discussion.

## 2. Some Trends

The speech field is making progress on a broad scale as demonstrated by the 900 or so papers and posters presented at the recent International Conference on Spoken Language Processing (ICSLP) in Beijing, October 2000. [To be illustrated by listing topics.] Three points may be made on the preceding list of current topics in speech research. Firstly, the wealth of topics that are being addressed in current fundamental and applied research obviously demonstrates that "the speech problem" has *not* been solved but continues to pose a series of major research challenges. [Mention some of them.] Secondly, the breadth of the speech topics that are being addressed could be taken as evidence that the speech field is simply doing *business as usual,* albeit on a larger and more ambitious scale than ever before. Thirdly, however, it is clear from the topics list that *the speech field is no longer separate from many other fields* of research but is in a process of merging into something which might perhaps be called the general field of interactive technologies. This latter trend, it may be argued, is the single most important factor which will influence the speech field in the future and which already suggests that the field is in a state of profound transformation.

**Interactive technologies**

It is relatively straightforward to explain why the speech field is gradually merging into the general field of interactive technologies. Since speech now works for a broad range of application purposes, a rapidly growing fraction of the speech research community are becoming involved in advanced interactive *systems* research rather than continuing to work on improving the speech *components* which form part of those systems. In advanced interactive systems research, speech is increasingly being used not as a stand-alone interactive modality as in, e.g., spoken language dialogue systems over the telephone, speech dictation systems, or text-to-speech systems, but as a modality for exchanging information with computer systems in combination with other modalities of information representation and exchange. Moreover, speech is not just an interactive technology among many others. Spontaneous speech is an extremely powerful input/output modality for interacting with computer systems, a modality which, furthermore, is available and natural to the large majority of users without any need for training in using it for interactive purposes.

The ongoing shift from speech components research to research on integrating speech in complex interactive systems has a number of important implications for the speech field. Speech researchers are becoming systems researchers and engineers. Far more than components research, systems research and engineering is exposed to the full complexity of today's world of information and telecommunications technologies. Few, if any, groups can build full systems on their own from scratch. To stay competitive, they have to *follow closely* the global developments in relevant systems architectures, platforms, toolkits, available components of many different kinds, de facto standards, work in standards committees, market trends etc. They need larger and much more *interdisciplinary teams* in order to keep up with competitive developments. They need *access* to platforms and component technologies in order to avoid having to do everything by themselves. And they need expertise in *software systems engineering best practice as specialised to the kind of systems*

*they are building,* including expertise in systems and usability evaluation. As we shall see in Section 4, they need even more than this, such as *hardware* access or expertise, development *resources, behavioural research* in new domains, and skills in *form and contents design.*

Compared to traditional research on improving a particular speech component technology, the world of advanced interactive systems research would appear to be orders of magnitude more complex. Moreover, that world is quite diffuse for the time being. It does not have a single associated *research community,* being inhabited instead by researchers from most traditional ITC (Information Technologies and Telecommunications) research communities. The world of advanced interactive systems research does not have any clear *evolutionary direction,* being characterised rather through ever-changing terms of fashion, such as 'ubiquitous computing', 'things that think', 'wearable computing', 'the disappearing computer' or 'ambient intelligence'. Significantly, all or most of those terms tend to refer to combined hardware and software systems rather than to components, and none of them refer to the traditional communities in the ITC field, such as speech processing, natural language (text) processing, machine vision, robotics, computer graphics, neural networks, machine learning, or telecommunication networks. Indeed, most of our current stock of inspired and visionary terms for describing the future of interactive technologies tends to be rather vague with regard to the technologies which they include or, if any, exclude.

Rather than trying to clarify what might be meant by the terms of fashion mentioned above, it may be useful to look at two other developments in conceptualising the field of advanced interactive systems research of which speech research has begun to form a part. To be sure, the concepts to be discussed are expressed by fashion terms as well, but at least it would seem that those concepts are of a more systematic and theoretically stable nature at this point.

**Natural interactivity**

When being together, most humans interact through speech when they exchange information. The telephone allows them to use spoken interaction at a distance as well, and the function of the telephone will soon be shared, or even taken over, by computing systems. When humans interact through speech, it does not matter if they are just a twosome or if they are more than two together. Moreover, except when speaking over the telephone, speech is not their only modality for information exchange. Gesture, lip movements, facial expression, gaze, bodily posture, and object manipulation all contribute to adding information, however redundant, to the spoken message. Together with speech, those modalities constitute *full natural human-human communication.* Moving beyond current technologies, we envision not just a single human speaking on the telephone or to a (desktop) computer in order to get a particular task done. Rather, the vision is one in which multiple humans speak together whether or not they are in the same physical location whilst using the system as an increasingly equal partner in communication. The system mediates their communication when needed, understands full natural communication, and produces full natural communication itself, increasingly acting as its human counterparts in communication. In order to take this vision into account, it would seem timely to abandon the traditional model of interaction which is called 'human-computer interaction', and replace it with the more general model of *natural human-human-system interaction* (HHSI). Natural HHSI, it appears, it a necessary end-point of current research in speech technologies. Thus, natural interactivity may serve as an important, even if distant, guidepost for the role of speech research in the complex world of interactive systems research.

The received picture of the role of theory in engineering goes something like this. It is hardly ever possible to deduce from theory a complete specification of the artefact that would constitute an optimal solution to some engineering problem. The reason is that the complexity

of the problem space involved always exceeds the power of theory. On the other hand, without theory (of physics, chemistry, computation etc.), it would not have been possible to build many of the artefacts we use in our daily lives. Thus, theory has a necessary supporting function in engineering. This is clear in the case of natural interactivity. To achieve the ultimate goal of natural HHSI, we need far better theory than is available at present: about how humans behave during natural interaction, about the behavioural phenomena which are relevant to the development of fully natural interactive systems, about how these phenomena are interrelated, about how they should be encoded etc. We also need a novel theory of natural communication which can replace speech acts theory and discourse theory by taking the notion of a complete communicative act as its basic notion.

**Multimodality**

The trend towards multimodal interactive systems reflects the trend towards blending of traditional research communities noted above as well as the increasing role of speech in future interactive systems. Multimodal systems are systems which offer the user combinations of input/output modalities for (or ways of) exchanging information with computer systems. Given the naturalness and expressive power of speech, speech input and speech output have the potential for becoming key modalities in future interactive systems. However, compared to natural interactivity, our current understanding of multimodality is much less capable of providing guideposts for future advanced interactive systems research in general and research on multimodal systems which include speech modalities in particular. Much too little is known about how to create good modality combinations which include speech for a variety of interactive purposes. This topic has become an active field of research, however (Bernsen 1997a, Benoit et al. 2000, Bernsen 2001). Further progress in this field is likely to complement research on natural interactivity in providing guideposts for speech-related research in the complex world of advanced interactive systems. In fact, these two research directions are intertwined in so far as it remains an open issue for which application purposes technologies, such as, e.g., animated speaking characters might provide useful solutions.

**Medium-scale science**

The final trend to be mentioned is the trend towards medium-scale science in advanced interactive systems research. Increasingly, it is becoming evident that the standard 3/4/5-team, low-budget, 3-year isolated advanced systems research project is often an inefficient means of achieving significant research progress. In many projects, the participants share discouraging experiences, such as the following: even if small, the project is only able to start almost one year after its conception because of the administrative processing needed to release the funding for the project; when the project begins, the participants discover that their objectives have already been achieved elsewhere; the participants spend the first half of the project trying to identify the best platform to work from only to discover that they cannot get access to it; the participants spend half of the project building and putting together a low-quality version of the contextual technologies they need before they can start addressing their core research objectives; at the start of the project, the participants realise that it will take too long to produce the data resources they need, such as tagged corpora, and decide instead to work with sub-optimal resources which they can get for free; etc. One way to avoid, or reduce the number of, such experiences is to launch larger-scale concerted research efforts which have a better chance of moving beyond the state of the art. World-wide, experiments are currently underway on how to carry out such medium-scale science. In the US DARPA Communicator project which addresses spoken language and multimodal dialogue systems, for instance, all participants start from shared core technologies without having to build these themselves (http://fofoca.mitre. org/). In the German SmartKom project which addresses multimodal

communication systems, the budget is large enough for the participants to build and integrate the technologies needed (http://smartkom.dfki.de/start.html). In the European Intelligent Information Interfaces (i3, http://www.i3net.org/) and CLASS (http://www.class-tech.org/) initiatives, whilst the traditional 3-year small-scale project topology has been preserved, major efforts are being made to promote cross-project collaboration, synergy, and critical mass.

For reasons too obvious to mention, relatively small-scale research should continue to exist, of course. Still, the complexity of the world of advanced interactive systems research is not likely to go away. This raises the question of whether we need more medium-scale science and less small-scale science in order to make efficient use of the funds available for advanced interactive systems research. If this question is answered in the affirmative, the important issue becomes how best to do medium-scale science, i.e. which model(s) to adopt for the larger-scale research efforts to come.

## 3. Estimated Technology Timelines

This section attempts to estimate the time of first appearance of a broad selection of generic and/or landmark speech technologies including natural interactivity technologies and multimodal technologies involving speech. Some qualifications are necessary to the proper interpretation of the proposed predictions. Despite the numerous uncertainties involved in estimating technology progress, timelines, when properly estimated, qualified, and peer reviewed, do seem a useful means of conveying a field's expectations to the outside world and serving as a basis for actions to be undertaken to support research in the field.

**Qualifications**

(a) As in all timeline forecasts, there is some uncertainty in the forecasts below with respect to whether the technology is deployable or will in fact have been deployed in products at the suggested time. The claim for the figures below rather tend towards the *deployable* interpretation which is the one closest to the point of view of research. The *actual deployment* of a deployable technology is subject to an additional number of factors some of which are unpredictable, such as company technology exploitation strategies, pricing strategies, and the market forecasts at deployability time. Thus, several years may pass before some of the technologies below go from deployability to actually being used in mass products. This implies that one cannot from the estimations below construct scenarios for the Information Society in which people in general will be using the described technologies at the times indicated. In other words, the years below refer to "earliest opportunity" for actual deployment in what may be sometimes rather costly systems to be embraced by relatively few customers. Similarly, given the fact that there are thousands of languages in the world, it goes without saying that a technology has been established when it works in at least one of the top languages, a "top language" being defined as a language used by developers in the more affluent parts of the world.

(b) Another point related to (a) above is to do with underlying "production platforms". For many advanced, and still somewhat futuristic, speech and language -related systems, it is one thing to have produced a one-of-a-kind demonstrator system but quite another to have produced the system in a way which enables oneself or others to relatively quickly produce more-of-the-same systems in different application domains. An example is the so-called intelligent multimedia presentation systems which will be discussed in more detail in Section 4. Several examples exist, such as the German WIP system and corresponding systems from the USA. However, as long as we haven't solved the problem of how to produce this kind of system in a relatively quick and standardised way, intelligent multimedia presentation

systems are not going to be produced in numbers but will remain research landmarks. The timeline list below mostly avoids mentioning systems of this kind, assuming for the kinds of systems mentioned that the "production platform" issue has been solved to some reasonable extent at the time indicated.

(c) There is some, inevitable because of the brevity of the timeline entries, vagueness in what the described technologies can actually do.

(d) It is assumed that, after a certain point in time which could be, say, 2006, the distinction between technology use for the web and technology use for other purposes will have vanished.

(e) There is no assumption about *who* (which country, continent, etc.) will produce the described landmark results. However, given the virtually unlimited market opportunities for the technologies listed as a whole, it is expected that a consolidated technology timeline list will command keen interest among decision makers from industry and funding agencies.

(f) There is nothing about (software) agent technologies below. It is simply assumed that what is currently called software agent technologies will be needed to achieve the results described and will be available as needed.

(g) In principle, of course, any technology timeline list is subject to basic uncertainty due to the "if anything is done about it" –factor. If nothing will be done, nothing will happen, of course. However, most of the technologies listed below are being researched already and the rest will no doubt be investigated in due course. The uncertainty only attaches to who will get there first with respect to any given technology, who will produce the product winners, and how much effort will be invested in order to achieve those results before anybody else.

**Technology timelines**

Blue: PH. Red: JP. Green: SK. Purple: Hessen.

**Basic technologies**

| | |
|---|---|
| Hypotheses lattices, island parsing, spotting in all shapes and sizes for spoken dialogue | 2001 |
| Continuous speech recognisers in OSs for workstations in top languages | 2002 |
| Continuous speech recognisers in mobile devices (10000 words vocabulary) in top languages | 2003 |
| High quality competitive (with concatenated speech) formant speech synthesis in top languages | 2003 |
| Task-oriented spoken dialogue interpretation by plausibility in context and situation | 2003 |
| Generally usable cross-language text retrieval | 2003 |
| Multilingual authoring in limited domains by constructing conceptual representations | 2003 |
| Usable ontological lexicons for limited domains | 2003 |
| Usable translation systems for written dialogues (multilingual chatting) | 2003 |
| Useful speaker verification technology | 2004 |
| Seamless integration of spoken human/machine and human/human communication | 2004 |
| First on-line prosodic formant speech synthesis in top languages | 2004 |
| Simple task-oriented animated character spoken dialogue for the web | 2004 |
| Concept-to-speech synthesis | 2004 |
| Stylistically correct presentation of database content | 2004 |

| | |
|---|---|
| Superficial semantic processing based on ontological lexicons | 2004 |
| Max. 2000 words vocabulary task-oriented animated character dialogue for the web | 2005 |
| Prosodic formant speech synthesis replaces concatenated speech in top languages | 2005 |
| Full free linguistic generation (from concepts) | 2005 |
| Robust, general meta-communication for spoken dialogue systems | 2005 |
| Writer-independent handwriting recognition | 2005 |
| Learning at the semantic and dialogue levels in spoken dialogue systems | 2006 |
| Useful multiple-speaker meeting transcription systems | 2006 |
| Task-oriented fully natural animated characters (speech, lips, facial expression, gesture) output (only) | 2007 |
| Context sensitive summarization (responsive to user's specific needs) | 2007 |
| Answering questions by making logical inferences from database content | 2007 |
| Speech synthesis with several styles and emotions in top languages | 2008 |
| Continuous speech understanding in workstations with standard dictionaries (50000 words) in top languages | 2008 |
| Controlled languages with syntactic and semantic verification for specific domains | 2008 |
| Large coverage grammars with automatic acquisition for syntactic and semantic processing for limited applications | 2008 |
| Task-oriented fully natural speech, lips, facial expression, gesture input understanding and output generation | 2010 |

**Systems**

| | |
|---|---|
| First personalised spoken dialogue applications (book a personal service over the phone) | 2002 |
| Useful speech recognition-based language tutor | 2003 |
| Useful portable spoken sentence translation systems | 2003 |
| Useful broadcast transcription systems for information extraction | 2003 |
| First pro-active spoken dialogue with situation awareness | 2003 |
| Current spoken dialogue systems technology for the web (office, home) | 2004 |
| Satisfactory spoken car navigation systems | 2004 |
| Current spoken dialogue systems technology for the web (in cars) | 2005 |
| Useful special-purpose spoken sentence translation systems (portable, web etc.) | 2005 |
| High quality translation systems for limited domains with automatic acquisition | 2005 |
| Small-vocabulary (>1000 words) spoken conversational systems | 2005 |
| Medium-complexity (wrt. semantic items and their allowed combinations) task-oriented spoken dialogue systems | 2005 |
| Multiple-purpose personal assistants (spoken dialogue, animated characters) | 2006 |
| Task-oriented spoken translation systems for the web | 2006 |
| Useful speech summarisation systems in top languages | 2006 |
| Useful meeting summarisation systems | 2008 |
| Usable medium-vocabulary speech/text translation systems for all non-critical situations | 2010 |

| | |
|---|---|
| Medium-size vocabulary conversational systems | 2010 |
| **Tools, platforms, infrastructure** | |
| Standard tool for cross-level, cross-modality coding of natural interactivity data | 2002 |
| Infrastructure for rapid porting of spoken dialogue systems to new domains | 2003 |
| Platform for generating intelligent multimedia presentation systems with spoken interaction | 2005 |
| Science-based general portability of spoken dialogue systems across domains and tasks | 2006 |

Other problems which were strongly felt when producing the list above include: (i) the fact that there is plenty of continuity in technology development. "Continuity" may not be the right term because what happens is that what is later perceived as a new technological step forward is constituted by a large number of smaller steps none of which could be mentioned in a coarse-grained timeline exercise such as the one above. General speaker identification, robust speech recognition in hard-to-model noise conditions, "real" speaker-independent recognition (almost) no matter how badly people speak, or pronounce, some language, are all examples of minute-step progress. (ii) Another problem is to do with speech in fancy-termed circumstances, such as 'ambient intelligence' applications. It may be that there is a hard-core step of technological progress which is needed to achieve speech-related ambient intelligence but then again, may be there isn't. Maybe this is all a matter of using the timelined speech technologies above for a wide range of systems and purposes. Similarly, it is tempting to ask, for instance: "When will I have a speech-driven personal assistant?". But everything depends on what the personal assistant is supposed to be able to do. Some personal assistant technologies exist already. Thus, it does not seem possible to timeline the appearance of speech-driven personal assistants even if this might be attractive for the purpose of advertising the potential of speech technologies.

**How well is Europe doing?**

No attempt has been made, so far, to annotate the technology timelines with indications of how well, or how badly, European research is doing and hence how likely it is that a particular technology will be made deployable in Europe before anywhere else. In most of the timelined cases above, this would seem to depend primarily on the financial resources and research support mechanism which will be available to European research in the coming decade. In some cases, the US is presently ahead of Europe, such as with respect to continuous speech recognisers in workstations or broadcast transcription systems. In other cases, Europe has the lead, such as in building a standard tool for cross-level, cross-modality coding of natural interactivity data, continuous speech recognisers in mobile devices, advanced spoken dialogue systems, and spoken car navigation systems.

**Beyond 2010**

Beyond 2010 lie the dreams, such as unlimited-vocabulary spoken conversational systems, unlimited-vocabulary spoken translation systems, unlimited on-line generation of integrated natural speech, lips, facial expression and gesture communication, unlimited on-line understanding of natural speech, lips, facial expression and gesture communication by humans, summarisation-to-specification of any kind of communication, multimodal systems solutions on demand, and, of course, full natural interactive communication.

# 4. Implications of the Timelines

When analysing the implications of the timelines in Section 3, a number of uncertainties come up with respect to how the market for speech products will develop. At present, most speech products are being marketed by some 5-10 major companies world-wide. These companies are growing fast as are hundreds of small start-up companies many of which use basic technologies from the larger technology providers. It may be assumed that this market structure will not continue in the future. Rather, speech recognition and synthesis technologies would seem likely to become cheap, or even free and open source, components which will come with all manner of software and hardware systems. The implication is that all ITC providers who want to, will provide value-added speech products and that the basic speech technologies will not be dominated by a small number of large suppliers. Some important share of the speech market, including de facto standards in various areas, will probably be picked up by large custom software and mobile phone technology suppliers, such as Microsoft and Nokia, but that is likely to happen in any realistic scenario for the coming decade. The conclusion is that, during the coming decade, speech will be everywhere, in all sorts of products made by all sorts of companies. But will speech be everywhere in bulk? This raises a second uncertainty.

In one scenario, speech will be present in all or most ITC products by 2010, and speech will be popular and will be used as much as input keys, input buttons, and output graphics displays are being used today. In another scenario, however, speech uptake will be slow and arduous. Several reasons could be given for the latter scenario. Thus, (a) it may take quite some time before speech recognition is being perceived by users to be sufficiently robust to make users switch to speech where speech is better ideally. (b) It may take quite some time before the field and the market has sorted out when to use speech as a stand-alone modality and when to use speech in combination with other input/output modalities. If these two (a + b) take-up curves do not grow in any steep manner, speech may still be widespread by 2010, but speech will still not be as important an input/output modality as it is likely to become later on. For the time being, we would appear to have too little information to be able to decide between the two scenarios just discussed. There is simply not enough data available on user uptake of speech technologies to enable a rational decision to be made.

**Exploitation today**

Already today, there is a great exploitation potential for speech technologies because of the simple facts that (i) the technologies which already exist in a few top languages could be ported to hundreds of other languages, and (ii) the types of applications which already exist can be instantiated into numerous other applications of similar complexity. At this end of the speech technology spectrum, the emphasis is on flexible and versatile production platforms, quality products, and low-cost production rather than on research. This is particularly true of low-complexity over-the-phone spoken language dialogue information systems using continuous speech input. Users would seem to have adopted these systems to a reasonable extent already. The same degree of user acceptance does not appear to characterise the uptake of, e.g., spoken language dictation systems or simple spoken command systems for operating screen menus. Even if purchased by widely different groups of users, the former would appear to be used primarily by professionals, such as lawyers and medical doctors, and the latter hardly seems to be used at all. Also, text-to-speech systems for the disabled and increasingly for all users, do appear to have a significant exploitation potential already.

**Key technologies: speech-only**

The timelines in Section 3 highlight a series of key speech-only technologies which are still at the research stage, including:

- prosody in on-line speech synthesis;
- multi-speaker broadcast and meeting transcription;
- speech summarisation;
- speech translation; and
- conversational spoken dialogue.

*Prosody in on-line speech synthesis*

Prosody in on-line speech synthesis is probably important to the speed of take-up of speech technologies because users would appear likely to prefer prosodic speech output to non-prosodic speech output. However, there do not seem to exist firm estimates as to how much prosody matters. Reasonably clear and intelligible non-prosodic text-to-speech already exists for some top languages and might turn out to be satisfactory for most applications in the short-to-medium term.

*Multi-speaker broadcast and meeting transcription*

Multi-speaker broadcast transcription forms the topic of massive US-initiated research at the moment and appears likely to start becoming widely used in practice relatively soon. Like *meeting transcription* technology, multi-speaker broadcast transcription technology has a large potential for practical application as well as for acting as a driving force in speech and natural language (text) processing research. Once multi-speaker broadcast speech audio and meeting speech audio can be useably transcribed so that first application paradigms for these technologies have been achieved, the transcriptions can be further processed by other technologies, such as speech summarisation and speech translation technologies. It would be very valuable for European speech research if Europe could launch a meeting transcription technology evaluation campaign before the US (evaluation campaigns will be discussed below).

*Speech summarisation*

Speech summarisation is being experimented with already, often by using text or transcribed speech instead of raw speech data. Speech and text summarisation technology including intelligent speech and text search would seem to hold enormous potential by enabling users to obtain at-a-glance information on the contents of large repositories of information. The same applies to related technologies, such as question-answer systems which enable the user to obtain answers to specific questions from large repositories of information. Progress in these fields is difficult because of the difficulty of the research which remains to be done. However, the difficulties ahead are counter-balanced by expectations that far-less-than-perfect solutions could help to establish first application paradigms which, in their turn, might help accelerate progress.

*Speech translation*

Despite the embattled 40-year history of language (text) translation systems, speech translation is now being researched across the world because of the realisation that far-less-than-perfect paragraph-by-paragraph translation could yield useful applications in the shorter term. In their turn, those first application paradigms could serve as drivers of further progress. The German Verbmobil project (http://verbmobil.dfki.de/), for instance, demonstrated just how difficult human-human spoken dialogue translation is. Once application paradigms have

been achieved, however, speech translation technology would appear set to gain an enormous market. Still, it may take quite some time before there is a massive growth in the market for speech translation products, due to the difficulty of the research which remains to be done.

*Conversational spoken dialogue*

For some time, the term 'conversational spoken dialogue' has been a catch-all for next-step spoken language dialogue systems, such as those explored in the DARPA Communicator project. However, the DARPA Communicator agenda remains focused on task-oriented dialogue, such as flight ticket reservation. Even if conducted through mixed initiative spoken dialogue in which the human and the machine exchange dialogue initiative in the course of their dialogue about the task, task-oriented spoken dialogue might not qualify as conversational spoken dialogue. Conversational spoken dialogue is mixed-initiative, to be sure, but in conversational spoken dialogue there is no single task and no limited number of distinct tasks which have to be accomplished. Rather, spoken conversation systems may be characterised as *topic-oriented.* It is the breadth and complexity of the topic(s) on which the system is able to conduct conversation which determine its strength. Research on spoken conversation systems is still limited. Obviously, however, spoken conversation systems hold an enormous application potential because they represent the ultimate generalisation of the qualities which everybody seem to appreciate in task-oriented mixed initiative spoken language dialogue systems.

**Key technologies: multimodal systems**

In addition to speech-only technologies, the timelines in Section 3 highlight a series of multimodal speech systems technologies which are still at the research stage in most cases, including:

- intelligent multimodal information presentation including speech;
- natural interactivity;
- immersive virtual reality and augmented reality.

*Intelligent multimodal information presentation including speech*

Intelligent multimodal information presentation including speech is a mixed bag of complex technologies which do not seem to have any clear research direction at the present time. The reason is that the term *multimodality,* as pointed out in Section 2 above, refers to a virtually unlimited space of combinations of (unimodal) modalities. Thus, Modality Theory (Bernsen 1997b, 2001) has identified an exhaustive developers' toolbox of unimodal input/output modalities in the media of graphics (or vision), acoustics (or hearing), and haptics (or touch) consisting of more than a hundred unimodal modalities. The number of possible combinations of these unimodal input/output modalities is evidently staggering and, so far, at least, no way has been found to systematically generate a subset of good and useful modality combinations which could be recommended to system developers. The best current approach is to list modality combinations which have been found useful already in experimental or development practice. Obviously, given the limited exploration of the space of possible modality combinations which has taken place so far, those combinations constitute but a tiny fraction of the modality combinations which eventually will be used in HHSI. The same lack of systematicity applies to the subset of useful modality combinations which include speech output and/or speech input. Thus, for instance, it is known that speech and static graphics image output is a useful modality combination for some purposes and that the same holds for combined speech and pen input into various output domains as well as for speech and pointing gesture input into, e.g., a static graphics map output domain. The qualifying term *intelligent* is being used to distinguish intelligent multimodal information presentation

systems from traditional multimedia presentations. In traditional multimedia presentations, the user uses keyboard and mouse (or similar devices) to navigate among a fixed set of output options all of which have been incorporated into the system at design-time. In intelligent multimodal information presentation systems, the system itself generates intelligent multimodal output at run-time. This may happen through run-time language and/or speech generation coordinated with run-time graphics image generation and in many other ways as well. Some years ago, a reference model for intelligent multimodal information presentation systems was proposed by an international consortium of developers (Computer Standards and Interfaces 18, 6-7, 1997). Since then, little systematic development has happened, it appears, which is probably due to the fact that the field is as open-ended at it is. Still, it would appear that (i) the field of intelligent multimodal information presentation systems is an extremely promising approach to complex interactive information presentation, such as in interactive systems for instruction tasks for which several output modalities are needed, including speech. In order to advance research in this field, research is needed on Modality Theory in order to identify potentially useful modality combinations as well as on next-step architectures and platforms for intelligent multimodal information presentation.

*Natural interactivity*

As argued in Section 2, fully natural interactive systems represent a necessary vision for a large part of the field of interactive systems. Furthermore, spontaneous speech input/output is fundamental to natural interactive systems. Given this (latter) fact, it would seem that speech research is set to take the leading role in the development of increasingly natural interactive systems. Already today, this research and development process can be broken down into a comprehensive, semi-ordered agenda of research steps. The steps include, at least, (i) *fundamental research on human communicative behaviour,* including identification of the relevant phenomena which are being coordinated in human behaviour across abstraction levels and modalities, such as speech prosody and facial expression; validated coding schemes for these phenomena; and standard tools for coding the phenomena in order to create research and training resources in an efficient and re-usable fashion; (ii) *speech and graphics integration* in order to achieve full run-time coordination of spoken output with lip movement, facial expression, gaze, gesture and hand manipulation, and bodily posture; (iii) *speech and machine vision integration* in order to enable the system to carry out run-time understanding of spoken input in combination with lip movement, facial expression, gaze, gesture and hand manipulation, and bodily posture; and (iv) *conversational spoken dialogue* as discussed above. Other relevant technologies include, i.a., machine learning and 3D graphics modelling of human behaviour. Although research in underway on (i) through (iv), there is no doubt that the field might benefit strongly from a focused effort which could connect the disparate research communities involved and set a stepwise agenda for achieving rapid progress. The application prospects are virtually unlimited, as witnessed by the consensus in the field that increased natural interaction tends to generate increased trust in HHSI.

*Immersive virtual reality and augmented reality*

It is perhaps less clear what are the speech technology application prospects of immersive virtual reality. Today, immersive virtual reality requires that users are wired up with 3D goggles, force feedback data gloves, data suits, and/or wired surfaces and other wired equipment, such as flight cockpits or bicycles. At the present time, it seems uncertain to which extent and for which purposes immersive virtual reality technologies will be found useful in the future. The primary purposes for which these technologies are being used to day are advanced technology exhibition and demonstration, and the building of rather expensive simulation setups, such as flight simulators. Furthermore, it is far from clear which role(s)

speech will come to play in immersive virtual environments. These remarks also apply to *augmented reality* technology.

**Other research and supporting measures needed**

In order to promote efficient research progress on advanced interactive systems which include speech as a modality, technology research is far from sufficient. As pointed out in Section 2, present and future advanced systems research takes place in an extremely complex context in which leading research efforts must incorporate global state-of-the-art developments in many different fields. World-leading speech-related systems research should be accompanied by the following kinds of research, at least:

- state-of-the-art generic platforms;
- generic architectures;
- hardware;
- specialised best practice in development and evaluation;
- standard re-usable resources;
- behavioural research;
- neural basis for human natural communicative behaviour;
- design of form and contents;
- porting technologies to languages, cultures and the web;
- the disabled;
- maintenance for uptake.

*State-of-the-art generic platforms*

In order to effectively aim at exploitable results from early on, speech-related systems research needs to build upon existing state-of-the-art generic platforms including APIs. If a state-of-the-art generic platform is not available to the researchers, either because it does not yet exist or because it is inaccessible for proprietary reasons, researchers have to build it themselves. This is not possible in small-scale research projects which have an additional research agenda which presupposes a working platform. The consequence is that the research project will either build upon some sub-optimal platform in order to complete the research agenda, or build a better platform but not complete the research agenda. Both consequences are unacceptable, of course, but the former may work temporarily if the research aims are very advanced ones. However, when the research aims have been achieved or, at least, somehow explored, there will typically be no practical way of continuing the research in order to produce a state-of-the-art generic platform which could bring the research results towards the market. Two implications seem to follow: (i) it would be highly desirable if companies could be encouraged to make their most advanced platforms accessible to researchers. (ii) If a state-of-the-art generic platform is missing altogether, it should either be produced in a separate project or projects should be made so large as to include platform development. Both implications would seem to require a transformation of existing European research funding mechanisms.

*Generic architectures*

It would seem likely that overall research speed and efficiency in Europe could be accelerated by research on *generic architectures* for future systems, such as conversational spoken dialogue systems, intelligent multimodal information presentation systems which include speech, or natural interactive systems. In the absence of research initiatives on generic

architectures for future systems, research projects are likely to specify idiosyncratic architectures which may satisfy their present needs but which do not sufficiently take into account global developments nor prepare for the next steps in advanced systems development. For the time being, there does not appear to be any European speech-related initiative in this field apart from the CLASS project which was launched in the autumn of 2000 (http://www.class-tech.org/). For efficiency, work on generic architectures should be done as a collaborative effort between many small-scale research projects and industry as in CLASS, or between a medium-scale research project and industry.

*Hardware*

Increasingly, advanced systems demonstrators require *hardware* design and development. For many research laboratories, this is a new challenge which they are ill-prepared to meet. Moreover, there is no strong tradition for involving hardware producers in the field of speech technologies, primarily because the need for involving them is a rather recent one. Ways must be found to forge links with leading hardware producers in order to make emerging hardware available to researchers. This problem has much in common with the platform issue discussed above.

*Specialised best practice in development and evaluation*

Advanced speech systems research is conducted in a software engineering space bounded by, on the one hand, general software engineering best development and evaluation practice and, on the other, emerging ISO standards and de facto standards imposed by global industrial competition. Between these boundaries lies software engineering best practice in development and evaluation specialised for various speech-related systems and component technologies. This field remains ill-described in the literature. Apart from the DISC project on best practice in the development and evaluation of spoken language dialogue systems (www.disc2.dk), some work on evaluation in EAGLES Working Groups during the 1990s (http://www.ilc.pi.cnr.it/ EAGLES96/home.html), various national evaluation campaigns, and planned work in CLASS, little work has been done in Europe. By contrast, massive work has been done on component evaluation in the US over the last fifteen years. The result is that the speech-related technology field is replete with trial and error, repetitions of mistakes, and generally sub-state-of-the-art approaches. These negative effects are multiplied by the presence in the field of a large number of developers who are new to the field.

Admittedly, the field of software engineering best practice in development and evaluation specialised for various speech systems and component technologies is difficult and costly to do something about under present conditions. Technology *evaluation* campaigns are costly to do and require serious logistics. Yet the US experience would seem to indicate that technology evaluation campaigns are worth the effort if carried out for key emerging technologies including some of the technologies described in this paper. When a technology has gone to the market, industry does not want to participate any more and rather wants, e.g., evaluation toolkits for internal use. For emerging technologies, however, technology evaluation campaigns are an efficient means of producing focused progress. In fact, all participants tend to become winners in the campaigns irrespective of their comparative scorings according to the metrics employed, because everybody involved learns how to improve, or when to discard, their technologies and approaches. For Europe, technology evaluation campaigns for key emerging technologies could be a means of creating lasting advances on its global competitors. In order to take care of the complex logistics needed for the campaigns, it is worth considering to establish a European agency similar to the US NIST (National Institute for Standards in Technology) whose comprehensive experience with technology evaluation campaigns makes it comparatively easy to plan and launch campaigns

in novel emerging technologies. Alternatively, NIST might be asked to undertake to run technology development and evaluation campaigns in Europe, provided that this does not offend political and industrial sensibilities too much.

Effective *development* best practice work specialised for speech technologies is difficult to do under the current European funding mechanisms. The reason is that development best practice work requires access to many different components, systems and approaches in order to create an effective environment for the discussion and identification of best practice. This environment can only be established across many different small-scale projects or within medium-scale projects. CLASS is the first example of such an environment.

*Standard re-usable resources*

The term *resources* covers raw data resources, annotated data resources, annotation schemes for data annotation, and annotation tools for efficient automatic, semi-automatic or manual annotation of data. Resources are crucial for many different purposes, such as research into coding schemes or the training of components. Also, resources tend to be costly to produce. This means that, if the relevant resources are not available, research projects often take the easy way out which is to use less relevant but existing and accessible resources for their research. The results are sub-optimal research results and slowed-down progress. Common to resources of any kind is the need for standardisation. If some resource is not up to the required standards, its production is often a waste of effort because the created resource cannot be used for anything useful. In its strategy paper from 1991, ELSNET (http://www.elsnet.org/) proposed the establishment of a European resources agency. This recommendation was adopted through the creation of ELRA (European Language Resources Agency http://www.icp.inpg.fr/ELRA/ home.html) in 1995. ELRA is now a world-recognised counterpart to the US LDC (Linguistic Data Consortium, http://www.ldc.upenn.edu/). Still, ELRA is far from having the capacity to produce on its own all the resources and standards needed for efficient research progress. By contrast with technology evaluation campaigns, Europe has been active in the resources area during the 1990s. Today, there is a strong need to continue activities in producing publicly available resources and standards for advanced natural language processing, natural interactive systems development, evaluation campaigns as described above, etc. Recently, the ISLE (International Standards for Language Engineering) Working Group on Natural Interactivity and Multimodality (http://www.isle.nis.sdu.dk) has launched cross-Atlantic collaboration in the field of resources for natural interactivity and multimodality.

*Behavioural research*

Humans are still far superior to current systems in all aspects of natural interactive communication. Furthermore, far too little is known about the natural interactive behaviour which future systems need to be able to reproduce as output or understand as input. There is a strong need for basic research into human natural communicative behaviour in order to chart the phenomena which future systems need to reproduce or understand. This research will immediately feed into the production of natural interactive resources for future systems and components development, as described above.

*Neural basis for human natural communicative behaviour*

Related to, but distinct from, basic research into human natural communicative behaviour is basic research into the neural basis for human natural communicative behaviour. In the heydays of cognitive science in the 1980s, many researchers anticipated steady progress in the collaboration between research on speech and language processing, on the one hand, and research into the neural machinery which produces human speech and language on the other. However, massive difficulties of access to how human natural communicative behaviour is

being produced by the brain turned out to prevent rapid progress in linking neuroscience with speech and language processing research. Today, however, due to the availability of technologies such as MR imaging and PET scanning, as well as the increasing sophistication of the research agenda for the speech technology field, the question arises if it might be timely to re-open the cognitive science agenda just described. Potential results include, among others, input to generic architecture development (cf. above), identification of biologically motivated units of processing, such as speech and lip movement coordination, and identification of biologically motivated modalities for information representation and exchange. Relevant research is already going on in the field of neuroscience but, so far, few links have been established to the fields of speech technologies and natural interactive systems more generally.

*Design of form and contents*

Yet another consequence of the increasing emphasis on systems as opposed to system components is the growing importance of form and contents design. It is a well-established fact that design and development for the web requires skills in contents design and contents expression which are significantly different from those which have been developed through centuries for text on paper. In order to develop good demonstrator systems for the web or otherwise, there is a need for strongly upgraded skills in the design and expression of multimodal digital contents. For instance, it is far from sufficient to have somehow gleaned that speech might be an appropriate modality for some intelligent multimodal information presentation instruction system and to have available a state-of-the-art development platform for building the system. To actually develop the system, professional expertise in form and contents design is required. At the present time, few groups or projects in the speech field are adequately staffed to meet this challenge.

*Porting technologies to languages, cultures and the web*

Right now, the gap between the "have" countries whose researchers have access to advanced speech and natural interactivity components and platforms, and the "have-not" countries whose researchers cannot use those technologies for their own purposes because they speak different languages and behave differently in natural interactive communication, seems to be increasing. There is therefore a need to *port advanced technologies to different languages and cultures* both in Europe and across the world. The market will close the gap eventually in its own way, of course. However, in order to rally the full European research potential in the field in a timely fashion, it would appear necessary to actively stimulate the porting of technologies to new languages and cultures. From a research point of view, the best way to make this happen might be to include in medium-to-large-scale projects the best researchers from "have-not" countries even if, by definition, those researchers have to spend significant time catching up on basic technologies and resources before being able to actively contributing to the research agenda.

There is another sense of the 'porting technologies' -phrase in which Europe as a whole risks falling behind global developments. It is that of *porting speech, multimodal and natural interactivity technologies to the web.* The claim here is not that this is not happening already. The claim is that this cannot happen fast enough. In order to increase the speed of porting technology to the web, it would seem necessary to strongly promote advanced components and systems development for the web. It is far from sufficient to wait until some non-speech technology has been marketed for the web, such as electronic commerce applications, and then try to "add speech" to the technology. A much more pro-active stance would appear advisable, including a strongly increased emphasis on form and contents design as argued above.

*The disabled*

Advanced technologies for the disabled have a tendency to lag behind technology development more generally for the simple reason that the potential markets for technologies for the disabled are less profitable. Correspondingly, advanced technologies development for the disabled tends to be supported by small separate funding programmes rather than being integrated into mainstream programme research. In many cases, however, it would appear that systems and components technologies could be developed for any particular group of users before being transferred into applications for many other user groups. To the extent that this is the case, there may be less of a reason to confine the development of technologies for the disabled to any particular research sub-programme.

*Maintenance for uptake*

Finally, the small-scale science paradigm of small and isolated research projects does not at all cater for the fact that, in the complex world of advanced systems research, a wealth of prototype systems, proto-standard resources, web-based specialised best practice guides, etc., are being produced which have nowhere to go at the end of the projects in which they were developed. Their chances of industrial uptake, re-use by industry and research, impact on their intended users, etc., might become very substantially increased if it were possible to maintain them and make them publicly accessible for, say, two years after the end of projects. For this to happen, there is a need for (i) a stable web portal which can host the results, such as the present HLT (Human Language Technologies) portal under development (http://www.HLTCentral.org); (ii) open source clauses in research contracts for technologies which have nowhere to go at the end of a project; and (iii) financial support for maintenance. These requirement are likely to impose considerable strain of current European research support mechanisms. However, with some legal effort and a modest amount of financial support, the many research results produced in the speech-related field in Europe which are not being taken up immediately and which are not within the remit of ELRA, could gain much more impact than is presently the case.

# 5. Proposed Actions

Early preparations for the European Commission's 6th Framework Programme (FP6) including IST (Information Society Technologies) research are now in progress. It is premature to make predictions with any degree of certainty as to how the IST part of FP6 will shape up. Current information suggests an increased emphasis on basic research compared to the present FP5. In addition, it is possible that FP6 will include opportunities for the medium-scale research initiatives which were called for on several occasions above, i.e. large-scale "clusters" of projects all addressing the same research topic in a coordinated fashion. Finally, the current covering title for FP6 IST research is "ambient intelligence" which is one of the terms of fashion quoted in the present paper. Given the timelines and their analysis above, it does not seem to matter much which covering term is being chosen for FP6. "Ambient intelligence" is as apt as several others for FP6 and future advanced interactive systems research but, as argued in Section 3, it is far from clear if ambient intelligence requires us to focus on any particular segment of future speech-related technologies. However, the possible, increased emphasis on basic research as well as the possibility of carrying out medium-scale science in speech-related technologies are to be welcomed in the light of the argument above.

*5.1 Research priorities for speech-related technologies 2000-2010*

Taking into our stride the transformations of the field of speech-related research from speech-only to interactive systems in general, and from components research to interactive systems research, the top priorities in speech-related technologies research are:

- multi-speaker meeting transcription development and evaluation campaigns;
- speech summarisation development and evaluation campaigns;
- speech translation prototypes, generic platforms, and generic architectures. Development and evaluation campaigns are highly desirable;
- conversational spoken dialogue prototypes, generic platforms, and generic architectures. Development and evaluation campaigns are highly desirable;
- next-step prototypes, generic platforms, and generic architectures for intelligent multimodal information presentation;
- next-step prototypes, generic platforms, and generic architectures for natural interactive systems.

As soon as theoretically and practically feasible, all of the above advanced speech, multimodal and natural interactivity technologies should be developed for the web including hardware, form and contents design. The fact that some top research priorities have been mentioned above emphatically does not preclude the desirability of continuing "business as usual" in the field of speech-related research, including continued research into *all* of the technologies which have been mentioned earlier in the present paper. On the contrary, business as usual is actually assumed by the above top priorities list which focuses on technologies over and above business as usual. This also applies to next-step research into already deployed speech-related technologies, such as mixed initiative, task-oriented spoken dialogue systems.

For basic research leading to novel concepts, theories and formalisations, the top priorities are:

- basic research into human natural communicative behaviour;
- a novel theory of natural communication which can replace speech acts theory and discourse theory by taking the notion of a complete communicative act as its basic notion;
- research on Modality Theory in order to identify potentially useful modality combinations;
- establishment of collaborative links to research into the neural basis for human natural communicative behaviour.

*5.2 Research organisation needed*

Medium-scale science is needed for, at least, the coordinated development of natural interactive systems prototypes, generic platforms, generic architectures, best practice in development and evaluation, and standard resources. A large, medium-scale science project with these objectives should include the porting of technologies to new languages and cultures.

It is quite possible that the medium-scale science model could be applied to research into other speech-related technologies, such as speech translation technologies, conversational spoken dialogue systems, or speech technologies for ambient intelligence.

For researchers in small-scale speech-related projects, in particular, the creation of a generic platforms and hardware "bourse" through contributions from European industry would be of great importance.

Finally, we should stop having research programme ghettos for technologies for the disabled.

*5.3 Infrastructural actions needed*

In order to promote maximum uptake of the research results produced, it would be highly desirable to have funding for low-cost ways of maintaining research results for later uptake.

Given the emphasis on technology development and evaluation campaigns above, Europe needs to establish an evaluation and standards agency. It is not evident to the present author that current political and industrial sensibilities would allow the US NIST to undertake to run technology development and evaluation campaigns in Europe.

This having been said, there is much to be said for increasing global collaboration on many aspects of speech-related research, such as creating a coordinated global infrastructure for resources distribution.

## References

Benoit, C., Martin, J. C., Pelachaud, C., Schomaker, L., and Suhm, B.: Audio-Visual and Multimodal Speech-Based Systems. In D. Gibbon, I. Mertens and R. Moore (Eds.): *Handbook of Multimodal and Spoken Dialogue Systems.* Dordrecht: Kluwer Academic Publishers 2000, 102-203.

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

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

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 2001 (to appear).

CLASS: http://www.class-tech.org/

*Computer Standards and Interfaces,* Special Double Issue, 18, 6-7, 1997.

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

DISC www.disc2.dk

EAGLES: http://www.ilc.pi.cnr.it/EAGLES96/home.html

ELRA: http://www.icp.inpg.fr/ELRA/home.html

ELSNET http://www.elsnet.org/

i3: http://www.i3net.org/

ISLE: http://www.ilc.pi.cnr.it/EAGLES96/isle/ISLE_Home_Page.htm

ISLE Working Group on Natural Interactivity and Multimodality: http://www.isle.nis.sdu.dk

HLT portal: http://www.HLTCentral.org

LDC http://www.ldc.upenn.edu/

SmartKom: http://smartkom.dfki.de/start.html

Verbmobil: http://verbmobil.dfki.de/