Modelling Spoken Multimodal Instructional Systems

Niels Ole Bernsen
Natural Interactive Systems Laboratory
University of Southern Denmark, Odense
Phone: +45 65 50 35 44
Email: nob@nis.sdu.dk

Laila Dybkjær
Natural Interactive Systems Laboratory
University of Southern Denmark, Odense
Phone: +45 65 50 35 53
Email: laila@nis.sdu.dk

Modelling Spoken Multimodal Instructional Systems

The use of speech and spoken dialogue is a relatively recent addition to instructional systems. As, almost invariably, human instructors and students *talk* during teaching and training, spoken dialogue would seem to be an important factor in systems that emulate aspects of human instruction. In this chapter, we describe the origins and state of the art of spoken multimodal instruction. We then discuss strengths and weaknesses of the speech modality, key roles of spoken dialogue in multimodal instruction, functional issues in current spoken teaching and training systems, commercial prospects, and some main challenges ahead.

1 Introduction

A key advantage of instructional systems is to enable instruction in the absence of a human expert or teacher. From pre-school kids to adults of all ages, everybody needs to learn and benefit from the expertise of others when doing unfamiliar tasks. The classical solution is to be helped by a *human instructor* who has two kinds of expertise: in the subject-matter in question and in effectively communicating or transferring the expertise to students. While this approach has worked for millennia, it suffers from the problem that expertise remains expensive and rare relative to the number of those who wish to acquire or draw upon it. A language instructor in class, for instance, has little time for coaching each student individually.

An interactive instructional system, or *system instructor*, offers to supplement the human instructor's contributions to individual student learning and problem-solving. In the ideal case, the system's expertise, both subject-wise and pedagogically, is near-equivalent to that of a good human instructor. Since systems can be copied infinitely, this would enable students to work with an expert all the time, in class, at home and elsewhere, and not just when the student has a human instructor's undivided attention in class. It is hardly controversial that removing the difficulty of access to expertise and dramatically reducing its price, is a worthwhile technological goal.

Below, we describe and discuss the roles of speech, spoken dialogue and conversation in instructional systems most of which include modalities other than speech. Characteristically, human instruction involves spoken conversation with students no matter whether spoken interaction is central to the instructional task or has an auxiliary role. In relative terms, speech is a newcomer in the field of instructional systems which for a long time was characterised by typed text input/output. Spoken interaction is insufficient for most instructional purposes, however. Other interactive modalities are needed for optimising instructional effectiveness and efficiency. New modalities and modality combinations hold the additional promise of providing system instructors for *all* users no matter their perceptual or motor disabilities.

We define instructional systems (Section 2), review their history and describe the state of the art of spoken instructional systems (3), and present conceptual architectures and component technologies (4). Using a simple example, we discuss how to approach instructional systems analysis and specification (5) and sketch a functional model of instructional interaction (6). Since speech is not a catch-all for instruction, we ask when (not) to use speech and propose key roles of spoken dialogue (7). We discuss examples of spoken multimodal dialogue systems (8), commercial prospects (9), and present some main research challenges (10).

2 Instructional Systems

By an (interactive) *instructional system* we understand an application whose main purpose is to teach or train the user or help the user solve a particular problem. Although often combined in practical applications, these goals are somewhat different. A *teaching system* primarily teaches *understanding* of some subject-matter, such as the periodic system, basics of genetics, astronomy, planet geography, phases in the history of humanity, etc. A *training system* primarily trains *practical skills*, such as language skills, how to operate some artefact, play golf, or fly a commercial airliner. Teaching and training systems are aimed at long-term learning effects in the learner. By contrast, *problem-solving support systems*, such as one helping to install IP telephony on a laptop, rarely incorporate ambitions of producing long-term learning effects. If they help solve the problem at hand, they fulfil their purpose.

Aiming at long-term retention which largely depends on the amount of elaboration done on the education material, *teaching/training systems* typically focus on providing opportunity for solving, or otherwise addressing, as many and as different problems or issues as possible in the application domain. Key challenges in developing a good system are to make it *pose* the right challenges, *evaluate* the student's attempts to cope, *feed back* evaluations, *monitor* progress, *modify* challenge level depending on learning progress, and *stimulate* motivation to continue learning. *Problem-solving support systems* focus on system problem-solving because the user is challenged already and needs help. Problem-solving support systems thus partially reverse the roles described above, so that the user poses the challenge, evaluates the system's attempt to cope, and feeds back evaluations - but the system is still the expert.

Instructional systems need a *usable interface* for human-system interaction. In a sense, this is no different from other interactive systems, like word processors or spreadsheet packages. Arguably, however, usability requirements are particularly sharp for instructional systems: nothing is more de-motivating to self-instruction than a system you cannot find out how to use; students often work alone or in small groups, lacking the usual support from colleagues at work when something is amiss; and students typically need all system functionality rather than the <20% functionality most word processor users actually use.

3 History and State of the Art

3.1 Intelligent Tutoring Systems

Intelligent machines for educational purposes date back to Pressey's [1927] machine for multiple-choice tests. Computer-assisted training and teaching dates back to around 1960. While the first computer-assisted instruction (CAI) or computer-assisted training (CAT) systems were fairly simple, one source of progress was incorporation of AI-techniques in the 1970s. This led Sleeman and Brown [1982] to coin the term intelligent tutoring systems (ITSs) to distinguish the new AI-based systems from simpler CAI/CAT systems.

One of the first ITSs was the WHY teaching system [Stevens and Collins 1977] which tutors factors and causal relationships affecting rainfall. A later, well-known training system is Sherlock and its successor, Sherlock II, which tutor air force trainees in diagnosing and repairing electronic equipment [Lesgold et al. 1992a, 1992b]. These are just examples of the multitude of domains addressed by ITSs over the years.

3.2 Early Intelligent Tutoring System Interfaces

For many years, ITSs were basically GUI (Graphical User Interface) –based, using input from keyboard and mouse and output on screen. Screen output was to begin with typically static text and graphics followed more recently by dynamic output, such as video, animation and

virtual reality. Since human tutoring typically involves natural language interaction, GUI-based instruction also began to include typed student-system dialogue. This trend seems to have grown with advances that now enable rather sophisticated linguistic interaction.

3.3 Natural Language Interaction

Some early text-based dialogue systems are psychotherapist Eliza [Weizenbaum 1966] and SHRDLU [Winograd 1971], the latter enabling users to move blocks of different shape and colour around by using a vocabulary of about 50 words. Theoretical work on discourse and dialogue in the 1970s and 1980s [Grosz 1974, Allen 1979, Grosz and Sidner 1986] has played a major role in advancing natural language interfaces. Spoken interaction began to gather speed around 1990. In the 1990s, most spoken dialogue systems enabled users to accomplish some task, such as making a flight reservation [Bernsen et al. 1998] or checking bank information, but few systems were instructional. An example of the latter is the speech-only Circuit Fix-It Shop problem-solving support system [Smith 1991, Smith and Hip 1994]. The system helps debug an electric circuit and a main development goal was to model mixed-initiative dialogue. Research on spoken and multimodal interaction goes back at least to Bolt's [1980] system which combines spoken commands and pointing-gesture input.

3.4 Spoken Teaching and Training Systems

Spoken interaction made its way into ITSs in the late 1990s. For instance, Graesser et al. [2001, 2004] use talking-head output in their AutoTutor system but still rely on text input. AutoTutor teaches Newtonian qualitative physics and computer literacy. The Conning Officer Virtual Environment (COVE) system is for training Navy officers to become better ship drivers [Roberts 2000]. Interaction is via graphics output and speech, the system using short spoken exchanges to coach the learner during simulation. The shipboard damage control trainer [Clark et al 2001, 2005] also uses spoken interaction and graphics output. Students must contain the effects of fire, explosion and other critical onboard events, and receive spoken instruction and feedback. The system asks question on, e.g., what to do in a particular situation and which steps to take. The user answers via speech and/or pointing to part of the vessel displayed on-screen. Teaching system ITSPOKE [Litman and Silliman 2004] uses the WHY2-Atlas [VanLehn et al. 2002] text-based ITS as back-end. When given a problem in qualitative physics, the student types a natural-language essay answer. ITSPOKE analyses the answer and engages students in spoken dialogue to provide feedback, correct misconceptions and elicit more complete explanations.

Since spoken dialogue systems began to go multimodal in the late 1990s, several dialogue research projects have explored spoken multimodal interaction for teaching or training. Compared to mainstream ITSs, the resulting systems tend to focus less on pedagogical aspects and more on interaction. For instance, several training applications using spoken dialogue with virtual humans have been developed at the University of Southern California. One is a mission rehearsal system for training critical decision-making skills in small-unit US army leaders [Hill et al. 2003]. Another is a negotiation trainer for military personnel who need good negotiation skills when going to war zones [Traum et al. 2005]. Focusing on improving conversational abilities, these projects go beyond the strict task-orientation of most spoken dialogue systems, towards enabling a more open conversation within the domain. This is even more so in the European Hans Christian Andersen system for non-task-oriented conversation with the fairytale author about his life, person, and fairytales [Bernsen et al. 2004]. Aimed at the 10-18 year olds, the system combines education and entertainment.

The Collagen (COLLaborative AGENt) project (http://www.merl.com/projects/collagen/) [Rich et al. 2001], although fostered in the spoken multimodal tradition, goes a long way

towards merging with the ITS tradition and its pedagogical emphasis. Collagen introduced a platform for building mixed-initiative assistants for a wide range of applications and with considerable software re-use. Underlying the platform is shared-plan collaborative discourse theory. The platform has been used for, e.g., an agent that teaches how to operate a gas turbine and one which helps set up and program a video-cassette recorder [Rich et al. 2001]. To bridge to ITSs, domain-independent pedagogical agent Paco has been developed and used for teaching students how to operate gas turbine engines [Rickel et al. 2002]. Language training systems is another example of systems that draw on both traditions. The Colorado Literacy Tutor [Cole et al. 2003] is aimed at teaching students to read fluently and understand what they read. Talking animated head Baldi teaches vocabulary and grammar to autistic and hard-of-hearing children and helps them improve speech articulation and linguistic and phonological awareness [Massaro 2005].

3.5 Speech in Commercial Instructional Systems

While many commercial instructional systems include text-to-speech output, there are still rather few that include speech recognition. Spoken output is primarily used to speak some text aloud. Although any text may be read aloud, most commercial instructional software providers who stress the availability of spoken output, use it for some kind of language training, cf. below. Similarly, most commercial instructional systems that recognise speech are aimed at language training.

Text-to-speech output is, e.g., included in the reading, grammar, and vocabulary improvement programs from Merit Software (http://www.meritsoftware.com). A product from Kurzweil (http://www.kurzweiledu.com) aims to ease and enhance the reading, writing and learning experience of the visually impaired by speaking text aloud. Knowledge quiz software from Interactive Speech Solutions and Microsoft's Mobility Solutions for Education (http://getccq.com) ask questions in English while the question text is displayed in the application window. Spanish-speaking students may click a button to hear the question spoken in Spanish while still viewing the English text. Several systems from Caltrox (http://www.caltrox.com) use spoken output, including programs for learning multiplication tables, teaching kids to count, learn the alphabet or the spelling of words, and a text-to-speech program for training English pronunciation and vocabulary-building.

Several commercial pronunciation training systems use speech recognition, including Auralog's (http://www.auralog.com) Tell Me More and one from Protea Textware (http://www.proteatextware.com.au). These programs use the speech recogniser's recognition of the student's pronunciation of words and phrases as a basis for feedback on the pronunciation quality. We use pronunciation training as an example in Section 5. Spoken dictation systems are also being used as a help for dyslectic students.

4 Components and Architectures

In this section we describe some basic aspects of components and architectures for ITSs and spoken multimodal dialogue.

4.1 Core Components of Intelligent Tutoring Systems

A typical ITS includes the following abstract components:

- the *student model* (user model) collects, stores and updates information about the individual student for use by the teacher model. As a minimum, the model keeps track of how well the student performs over time;

- the *teacher model* (pedagogical model) is a model of the teaching process adaptable to the individual student's needs. The model includes, e.g., information about when to introduce a new learning topic depending on curriculum and student model information, and decides on the performance evaluation feedback to present to the student;
- the *expert model* (domain model) includes information relating to what is being taught as well as a model of how an expert solves problems in the domain. This enables comparison with the student's solutions and helps identify and point out which problems of understanding and/or skills mastery the student may have;
- the *user interface* presents learning material to the student and generally enables student-system interaction.

These functionalities can be realised in many different architectures and may vary hugely in sophistication.

4.2 Core Components of Spoken Dialogue Systems

A typical architecture for spoken natural language dialogue includes the following modules, cf. Figure 1:

- *speech recognition* transforms the speech signal into one or several text strings;
- *natural language understanding* extracts semantics from the recognised string(s);
- *dialogue management* decides, basis on input semantics and contextual information, which output to produce next;
- *natural language generation* prepares an output text string in accordance with the dialogue manager's decision;
- *speech synthesis* transforms the output text into a speech signal;
- *application data and business logic* provides backend information for the dialogue manager. Its contents depend on the task and domain addressed.

If interaction is text-based-only, speech recognition and synthesis are left out. If spoken interaction is not dialogue but only, e.g., uninterpreted single-word input/output as in the pronunciation trainer (Section 5), natural language understanding and generation are left out and spoken dialogue management reduced to more basic interaction management. Multimodal interaction requires additional input recognition and interpretation components and often also fusion of information received in different modalities, and/or additional output generation and rendering components, possibly including modality fission.

Space does not allow detailed discussion of spoken dialogue system component technologies and their pros and cons, see [McTear 2004, Delgado and Araki 2005]. Briefly, instructional systems developers may use commercial or research *speech recognisers* depending on the recognisers available for the language(s) used, quality requirements, and whether commercial recognisers offer the particular features needed, e.g., extraction of prosodic cues. In the large majority of practical systems, *natural language understanding* of spontaneous spoken or typed input is based on shallow (or robust) parsing which extracts key words, phrases, and possibly certain grammatical constructs from the input string for building a conceptual representation of the input, using analytical or statistical techniques and typically being guided by transcribed corpora of the kind of dialogue the system should be able to engage in. Deep parsing based on some comprehensive grammar fragment is generally not sufficiently robust vs. recognition errors, tends to get lost in multiple input interpretations, and is not needed to obtain usable results. In most cases, *dialogue management* must be developed from scratch unless the instructional system developers are on their way to having a platform which allows partial re-use from other systems. *Natural language generation* is typically based on stored output templates which are completed at run-time based on user input details and

possibly on system state properties as well. *Spoken output* may be pre-recorded human speech or – for increased flexibility - produced by free or commercial text-to-speech synthesisers which have achieved high levels of intelligibility and naturalness for several languages. Finally, application-specific data (teacher model, expert model) must be developed from scratch as must the student model unless the developers have partially reusable components from similar systems.

Summarising, the addition of spoken dialogue to instructional systems implies non-trivial investment in a family of technologies, most members of which are not simply off-the-shelf components. Even if one chooses off-the-shelf recognition and text-to-speech, and unless the application requires basic dialogue capabilities-only, one is into non-standardised software development of, and contents provision for, natural language understanding, dialogue management, and natural language generation.

4.3 Spoken Multimodal Instruction Architecture

Since ITSs are primarily characterised by the right-most backend components in Figure 1 while spoken multimodal systems are mainly characterised by the other components shown, combining the two parts produces an ITS with a spoken multimodal interface. Figure 1 simplifies input fusion which may be done at different input processing stages, primarily at signal and semantics level. In output fission, the output decided upon by interaction management is split between output modalities, such as speech and graphics, making sure that proper temporal synchronisation is maintained [Martin et al. 2006].

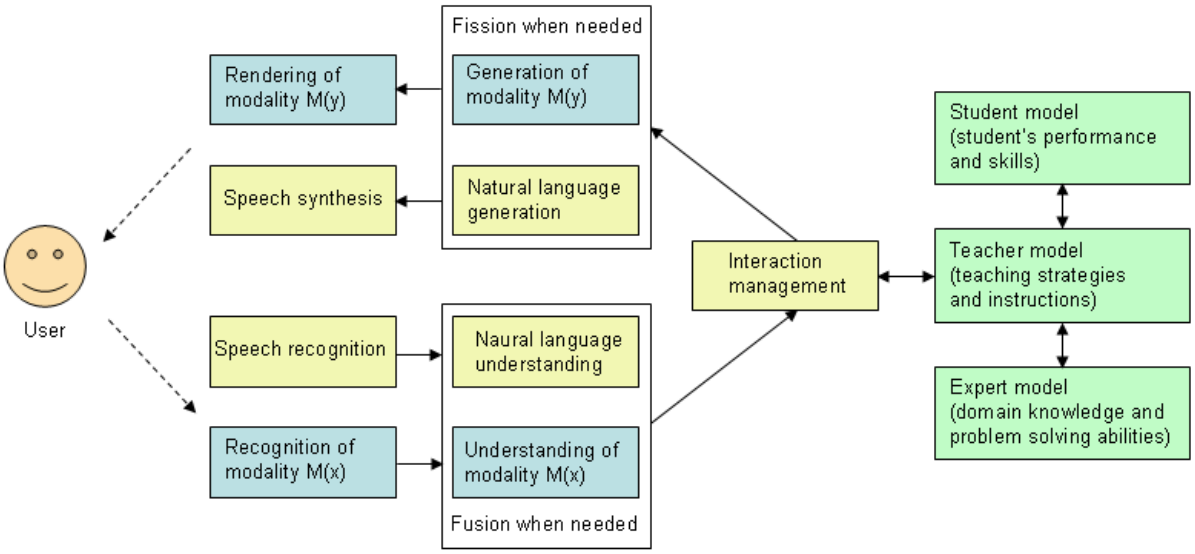

Figure 1. Conceptual architecture of a spoken multimodal instructional system.

5 System Analysis and Requirements: A Simple Example

Instructional system development follows general software engineering principles Sommerville [2006], adapting these to the application at hand. Decision on whether to use speech and spoken dialogue should be made early in the lifecycle as sketched for a simple spoken multimodal training system below. The example also illustrates basic student-, teacher-, expert- and user interface models in action. The system does not include spoken dialogue but might come to do so later.

Early lifecycle work focuses on analysing the target system and specifying requirements. For analysis and specification, we recommend consideration of a standard set of factors [Bernsen and Dybkjær 2007, to appear] which helps structure analysis and determine requirements:

(1) application type;
(2) user, i.e., general user properties to be taken into account;
(3) user profile, i.e., description of the target user group(s);
(4) use environment;
(5) domain;
(6) task or other activity;
(7) interaction;
(8) interaction device.

Assuming that our target application is a pronunciation trainer, let's sketch how those factors might influence its specification.

Even the most cursory target application description typically carries implications with respect to several factors. A description (1), such as "a speech recognition-based system for training immigrants in Danish single-word pronunciation" [Bernsen et al. 2006], implies the goal of improving student *skills* rather than knowledge and understanding (2) in the Danish second-language *domain* (5). Training should be in some quiet *use environment* (4), the *user profile* (3) being, if feasible, that of immigrants-in-general. Within the domain, we need a corpus of words which covers all Danish phonetic variations. Basically, the student's *task* (6) is to pronounce the words one-by-one. Decision on which *input/output modalities* to add to spoken single-word input, and in which interactive roles (7) - is less straightforward, which carries over to the choice of *interaction devices* (8). An open issue is how each training word should be presented to the student, e.g., via (a) typed text, (b) an audio file of a native speaker's pronunciation, (c) an audio/video file, (d) a semi-transparent animated human head pronouncing the word and displaying the vocal tract in action, or some or all of these. Another issue is whether to use spoken dialogue for some interactive role or if, e.g., a standard GUI environment is sufficient.

In the current pronunciation trainer version, the *expert model* includes the phonetically rich training vocabulary presented to students and algorithms for evaluating student pronunciation. The expert output is simply a set of pre-recorded text, audio and video files. An animated head is planned [Hansen 2006]. In the *student model*, we distinguish between basic and generalised results. *Basic results* are the logfiles stored each time a student pronounces a word and the system rates pronunciation quality based on phonemic similarity with a native speaker's pronunciation of the same word. *Generalised results* are computer over basic results, e.g., an accumulated numerical score for consecutive pronunciations of different words or identification of a set of pronunciation problems for a particular student. The *teacher model* is relatively simple. Since Danish single-word pronunciation has no levels of difficulty (it's all hard!), the teacher model essentially (i) gives feedback on each pronunciation and (ii) uses student model generalisations to present pronunciation problem diagnoses and suggest remedial training exercises. Finally, the relatively simple *user interface* requires user identification through id entry (ensuring that user modelling models the right user) and enables training word selection, optional word presentation(s), word pronunciation, pronunciation quality score presentations (individual and cumulative), diagnostic feedback and (planned) audio or video replay of the pronunciation (Figure 2).

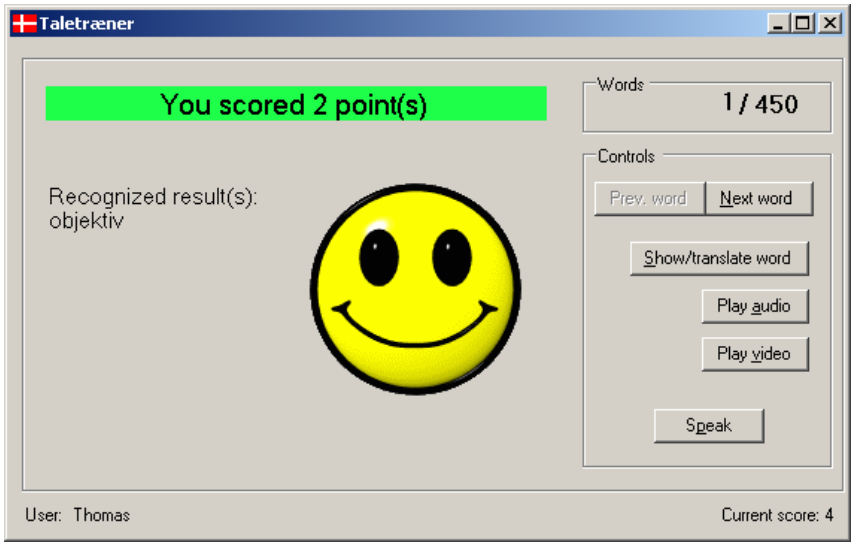

Figure 2. Pronunciation training system with max. score Smiley emoticon feedback.

Interestingly, the hardest part of pronunciation trainer development is one that tends to challenge many instructional systems, i.e., to produce optimal, pedagogically meaningful student model results and teacher model feedback on performance and progress. Simply put: it tends to be easier to *evaluate* student performance quality than to *diagnose* its deficiencies and propose tailored remedial action. This is where human teachers and trainers excel, partly because their perception of student performance details is keener, and partly because their reasoning about those details is sharper than what current systems can do.

6 A Model of Instructional Interaction

Spoken multimodal instruction is characterised by the fact that a particular modality, speech, is used in multimodal combination. To address the potential roles of speech and spoken dialogue in instructional systems, we need a model of instructional interaction. This section sketches a general model of a (self-) teaching or training session illustrated through reference to rather different system examples: the pronunciation trainer, case-based teaching of medical patient interviews, math training, the negotiation trainer and the Andersen system (Section 3.4), flight simulation, and systems for teaching models in physics and ecosystems.

6.1 Expert Knowledge

Large parts of system contents and behaviour can be fixed in advance because these are independent of the input users might produce.

The developers can fix (1) the *problem space* - the words to be pronounced, a set of medical interview cases or math problems, a set of negotiation goals or flight simulation targets, or a model that should be worked upon. These example systems are all task-oriented: the problem space is a *task space* in which the student works. Only the Andersen system has no task space because it is not task-oriented but *domain-oriented*. What the developers can fix is the person's personality, physical appearance, domain knowledge, habits, etc. (2) For task-oriented systems, developers can fix the nature and number of *actual problems* to be solved, such as pronouncing each word, critiquing each interview case, solving each math problem, reaching negotiation goals or simulation targets, or solving particular ecosystem problems. This cannot be done for the famous-person system because it is very much an empirical question what the students would want to learn from the person. (3) For each problem, the developers can determine the *solutions* that will be accepted as being correct and to which

extent. Correct solutions may be defined either in terms of following *correct procedure* or arriving at *correct results*, or both. The famous-person system has neither correct procedure nor correct results.

6.2 User Interface

The problem space should be presented using the modalities most appropriate for the purpose, such as text, speech, video and animated-face graphics for pronunciation training; static text for medical interview cases and static math notation for math problems, giving the student time to carefully study each problem; speech and animated conversational characters for negotiation and famous-person; a haptic-visual cockpit environment for flight simulation, possibly augmented with acoustic alarms and spoken controls; and some combination of static text and static/dynamic graphics for physics and ecosystem models. Some implications are that (i) *any* modality or modality combination might be useful for problem space representation in some particular instructional system; (ii) many, if not most, problem space representations are *inherently multimodal*; and (iii) only a fraction of problem space representations have *spoken dialogue* as main interactive modality, as in the negotiation and famous-person examples.

6.3 Student and Teacher

Let's add the student to the model. Typically, the system is a longer-term companion which the student (4) first needs to *learn how to use* and then uses for some period of time to improve knowledge or skills. Working with the system, the student must (5) *understand the problem space* as presented, (6) *understand the problem*s to solve, and (7) try to *solve the problems and present solutions.* The system must (8) *evaluate each solution* and feed back evaluations of process and/or results, (9) *generalise evaluations* of performance in order to spot patterns of difficulty in the student's problem-solving, *and present its generalisations* together with suggested remedies for removing the observed patterns of difficulty. Remedies might include special sessions for solving problems of a certain kind or increasing the level of difficulty for successful students. (9) requires observation of the individual student on-line and building of a model of that student's performance based on the observations made. This process is, in principle, the same for all input modalities: (mouse or hand) pointing, spontaneous speech, free text, etc.: the system evaluates inputs one-by-one, accumulates the evaluations, spots patterns, compares training/test results over time, updates the student model, etc., and uses the resulting information to guide learning based on recommendations from the teacher model.

6.4 Model Variations

Finally, let's add some model variations. Some systems might (10) distinguish between (a) *training/learning sessions* and (b) *test sessions*, e.g., making (a) more free-style with ample pedagogical feedback and (b) more formal with no pedagogical feedback but with performance scoring that enables easy comparison between test sessions over time. Some might (11) *reverse the roles of student and teacher/trainer*, the student teaching the system how to, e.g., solve equations or model an ecosystem, the system asking questions in the process [Biswas et al. 2005]. Another interesting perspective is (12) *multi-user* teaching/training systems where several students work together.

The model above needs not be fully implemented for a system to be instructional. Flight simulators, for instance, are typically just that, mission presentation, system operation instructions, and process and result feedback evaluations being provided by human

instructors. Rather, the model is an ideal model aimed to include all the functionality necessary for an instructional system to enable self-training unaided by human instruction.

7 When (not) to Use Speech in Instructional Systems

Section 5 listed factors to consider when specifying an instructional system. Section 6 described a model of the kinds of information to be exchanged at the user interface. This provides the background for asking when (not) to use speech and spoken dialogue in teaching and training systems, in which roles, and possibly combined with other input/output modalities. A *modality* is defined in modality theory as *a particular way* of representing information in some *physical medium*. Today, the three principal media used for interacting with computers, and the corresponding human senses, are: light/vision, acoustics/hearing and haptics (mechanical contact)/touch [Bernsen 2002].

7.1 Strengths and Weaknesses of the Speech Modality

Speech has several properties that make it well-suited as a main modality in instructional interaction, which is why speech is being widely used by human instructors. Studies of speech in multimodal contexts show that these properties are, among others described in [Bernsen 1997, Bernsen and Dybkjær 1999]:

- speech is a natural human modality for (1) *situated discourse* in which situation- or context-dependent information is being exchanged rapidly and spontaneously between interlocutors, each of whom can take the initiative;
- speech, and language more generally, has (2) *very high expressive potential*, so that virtually any piece of information could, in principle, be expressed in speech;
- compared to written language that evolved for non-situated information exchange, speech is more expressive due to (3) the *richness of the acoustic signal* which conveys far more than linguistic content, including emphasis, emotion, attitude, urgency, etc.;
- an acoustic modality, and unlike graphics and haptics, speech is (4) *omni-directional*;
- cognitively, speech can be effectively understood and generated in most (5) *heads-up, hands-occupied* situations, such as in the flight simulator;
- speech has (6) *high saliency*, i.e., is quite attention-catching.

On the negative side, the high salience of speech (6) can become a source of distraction both for the student and others. Moreover,

- speech, being (7) *temporal and transient*, does not offer the advantages of static modalities, such as static graphics or haptic text, of allowing students free perceptual inspection of the information conveyed. That's why we found that users prefer static typed text over speech when exchanging exact, high-complexity information and discussion summaries, whereas their discussions were all spoken conversation [Bernsen and Dybkjær 2001];
- speech is ill-suited for expressing (8) *highly specific and detailed spatial information* like the contents of images and spatial 3D scenes, or exact spatial locations. This is why many instructional systems use static and dynamic *output* graphics, such as images, data graphics, video, or virtual and augmented reality information presentation - to complement speech or otherwise. For *input*, this is why it is useful to combine pointing gesture and speech to enable users to point to objects and events instead of trying to explain their locations in speech. For similar reasons, speech input is mostly a poor replacement for (haptic) object manipulation by hand;

- speech input and/or output must be replaced by other modalities, e.g., sign language or written text, if users are (9) *hard-of-hearing* or have *speech disabilities*.

These speech modality properties contribute towards explaining why the problem space of instructional systems is often dominated by non-speech modalities. Among our benchmark systems, the pronunciation trainer combines speech input with "canned" output in a GUI environment; the medical patient interview and maths systems focus on static written text (property 7 above); the negotiation training and Andersen systems have spoken conversation at centre-stage together with animated interface agents as both systems explore situated human discourse and the latter explores combined speech/pointing gesture as well (properties 1, 3, 8); the flight simulator space is dominated by haptic input control and augmented-reality vision but has an auxiliary role for speech (4, 5, 6, 8); and the model teaching system spaces are dominated by various forms of text and output graphics (7, 8).

Increasingly, spoken dialogue applications include output *talking faces* or *embodied animated characters* [Cassell et al. 2000]. From one point of view, this is natural because human speech forms part of comprehensive communicative acts which include facial expression, gaze, gesture, and more; from another, some argue that the animations contribute little to instructional interaction and waste screen real-estate better used for presenting instructional task-related information; in a third view, their presence adds to instructional interaction more of the personal and expressive aspects characterising human instruction. This is an ongoing debate [Ruttkay and Pelachaud 2004]. In some cases, the animated face or agent is key to the application, e.g., the semi-transparent human head demonstrating vocal tract articulation for pronunciation training; when the embodied character acts as physical training instructor; or when students learn from conversation with a life-like person from another age.

7.2 Roles of Spoken Dialogue in Instructional Systems

What are the most important roles of spoken dialogue or conversation in problem-solving-oriented instructional contexts? Given the enormous diversity of potential applications, target user populations, etc., and the limited number of spoken multimodal instructional systems developed so far, the best approach may be to learn from the roles of spoken dialogue in human instruction.

Let us refine the question by considering a limiting case. When a human instructor is present and speech is not replaced by alternative options for situated discourse, e.g., sign language, spoken teacher-student dialogue becomes possible. Given the expressivity of spoken dialogue, teacher and student will almost unavoidably talk from time to time, asking, discussing, clarifying, helping, etc. However, we can define a limiting case in which spoken dialogue is unnecessary, i.e., when the instructional task is completely *self-explanatory* to the target users. A system that comes close is the Baldi language tutor which uses Baldi's talking face to improve the vocabulary of autistic and hard-of-hearing children [Massaro 2005]. Roughly, Baldi shows a screenful of, say, vegetables in static graphics; names them; asks the child to click on, e.g., the zucchini; praises or gives another try; asks about another static image, etc.; and moves on to a new screen. Baldi actually speaks to the kids and so might a human instructor, but the point is that everything is so straightforward that spoken *dialogue* is rarely needed.

Taking self-explanatory instruction as a limiting case, we propose that spoken dialogue may be added to instructional interaction for three main purposes, i.e.:

- task-oriented dialogue about the core teaching or training tasks, i.e., when the student solves a problem in dialogue with the system, including system feedback on the problem-solving process or solution;

- non-task-oriented conversation, primarily when the problem space itself is one of complex spoken dialogue or conversation but also when, e.g., other tasks and solutions are less clear-cut and require discussion;
- meta-communication about the interaction, including handling of miscommunication, help dialogue, introductory dialogue about instruction purpose, problem space, how to use the system, etc.

8 Spoken Dialogue in Multimodal Instructional Systems

Section 7 identified three main roles for spoken instructional dialogue. Based on these, we now illustrate and discuss how far we are in exploiting spoken dialogue for multimodal instruction. Task-oriented dialogue and task-transcending conversation are discussed in Section 8.1, meta-communication in Section 8.2, and spoken interaction and learning gain in Section 8.3.

8.1 Spoken Dialogue in Task-Solving and for Conversation

Today, nearly all instructional systems that include spoken dialogue are task-oriented and use limited *mixed-initiative* dialogue. Fully *user-directed* dialogue (no system initiative) seems unsuited for instruction while purely *system-directed* dialogue prohibits any kind of spoken intervention by the student. Limited mixed-initiative is typically obtained by carefully crafting the system's dialogue as illustrated in Figures 3 through 8. An obvious question is why these systems don't allow free mixed-initiative like in human instruction. Some reasons are:

- speech recognition errors;
- vocabulary and (semantic) grammar; and
- task and domain delimitation.

Paco: So, we succeeded in stopping engine two.
Student: Ok.
Paco: You take it from here.
Student opens fuel valve one on engine one.
Paco: Good.
Student opens the bleed valve on engine one.
Paco: That was a correct action, but I wasn't expecting you to do it now.
Paco: Let's go back to opening the fuel valves on engine one.
Student opens fuel valve two on engine one.
Paco: Great.
Student presses the engage button on engine one.
Paco: Whoops, you can't do that yet.
Student: What next?
Paco (Pointing): Press the on button on engine one.
Student presses the on button on engine one.
Paco: Good.
Student: I think I should set the throttle speed.
Paco: Right.
Student: What should the speed be?
Paco: The speed should be stop.

Figure 3. Example from [Rickel et al. 2002].

Tutor: Since you had some difficulty ordering boundaries around compartments, let's work through some drills.
Tutor: I will pick a couple compartments, and you answer the questions about boundaries for each one.
Tutor: Let's start some drills with primary boundaries, and then move on to the secondary boundaries.
Tutor: Let's assume there is a fire here, in 3-220-2-Q.
Tutor: What are the primary bulkheads you should use for setting boundaries on compartment 3-220-2-Q?
Student: Primary forward two two zero primary aft two five four
Tutor: Great job. Those are both right.

Figure 4. Example from [Peters et al. 2004].

Spontaneous speech recognition is error-prone, for several reasons. One is lengthy input: the longer the input the more likely misrecognition becomes. Since free initiative may encourage long input, ways must be found to reduce input length to keep misrecognition tolerably low. This can be done by carefully crafting the system's output in order to control initiative and limit input length. The alternative is to impose use of fixed spoken keywords and phrases but this increases student learning overhead and is infeasible for all but quite small-vocabulary input.

Systems allowing spontaneous speech input are often challenged as regards the sufficiency of their vocabulary and grammar. The solution is to continuously collect user input data and bootstrap the system on this data until it performs satisfactorily. This may require considerable effort. The more open the domain is, the more data is needed. Clear and clearly communicated task and domain delimitation is crucial, so that the student knows exactly what the system can and cannot conduct dialogue about. Inappropriate delimitation runs a high risk that users address out-of-domain issues, creating recognition and grammar problems [Bernsen et al. 1998]. The system's last defence is meta-communication which is difficult to engineer and always disturbs dialogue smoothness (Section 8.2).

ITSPOKE: I didn't understand. What force acts on all objects within close proximity to earth?

Student: uh the force of gravity

ITSPOKE: Fine. Besides the Earth's gravitational force, are there any other forces acting on the person, keys, and elevator after the release?

Student: no

ITSPOKE: Yep. Thus, keep in mind that after the release, the only force on the person, keys, and elevator is the force of gravity. If gravity is the only force acting on an object, will it be moving or staying still?

Student: it will be moving

Figure 5. Example from [Litman and Silliman 2004].

Tutor: So, when you add forces together, besides the magnitudes of the forces, what must you also consider?

Student: direct- direction (ASR: the truck in direction) [correct, uncertain, non-frustrated/ angry]

Tutor: Yes. [...] The vector sum of the forces acting on the body is also equal to the mass of the body multiplied by its acceleration. What is another term for the vector sum of forces acting on a body?

Student: net force (ASR: net force) [correct, neutral, non-frustrated/angry]

Tutor: That's super. Another term for the resultant force is the net force. In this relationship, what is the direction of the net force applied to a body and the direction of the body's acceleration?

Student: same same same (ASR: seems same type seems) [correct, certain, frustrated/ angry]

Figure 6. Example from [Forbes-Riley and Litman 2006]. Hand-annotated emotions in square brackets.

Figures 3 through 8 show instructional dialogue engineering aimed at minimising interaction problems. System questions are typically closed either through explicitly listing the answers to choose among (e.g., last tutor output in Figure 5) or by inviting short specific answers (e.g., first tutor questions in Figures 4 and 5, second in Figure 6). If the output invites student *initiative*, the tacit assumption is that the system will remain in control. Thus, in Figure 3, Paco offers initiative with little risk because the answer involves haptic screen graphics action rather than complex oral explanation. The Figure 7-dialogue is higher-risk because it opens up towards conversation. There is still a task to solve, however, i.e., to convince the doctor that he must move. Figure 8 goes further by showing real conversation with the fairytale author. Since

there is no instructional task, the developers can only control student input by making the character try to gently constrain the domains and topics of conversation.

Smooth and cooperative dialogue are major goals in most task-oriented applications, instructional or otherwise. Student cooperativity can be taken for granted as long as the student wishes to learn and manages to find out what (not) to talk to the system about. However, good human instructors are not merely cooperative in an extended Gricean sense of being to-the-point, taking into account student's background knowledge, etc. [Grice 1975, Bernsen et al. 1996] but re-phrase and re-explain when needed, never leave the student behind, tune the level of difficulty to the individual, and motivate and encourage even if the student hasn't got it yet. This is non-trivial pedagogical art even for human tutors. Evaluative feedback may be given in many different ways. It may be simple like the smiling or sad emoticon face for pronunciation training (Figure 2) which only reflects the quality of the most recent pronunciation; comment nicely on the most recent student action (Figure 3); or be general as in Figure 4 where the system explains why the student gets the next task. The Figure 7-system gives no explicit feedback on student performance. Rather, the performance measure is how willing to cooperate the doctor becomes.

C(aptain): Hello Doctor Perez.

D(octor): Hello.

C: I have orders to move this clinic to another location.

D: You want to move the clinic?

C: Yes

DECREASES SOLIDARITY: captain endorses undesired act 'run-clinic-there'

D: Look at these people!

D: we need to help them.

C: It is not safe here, we can't protect you.

DECREASES CREDIBILITY: captain asserted unbelieved (but possible) state 'patients-unsafe-here' 'patients-unsafe-here' could be established by captain's act of 'planned-attack'

D: Protect me? Protect me from what?

D: Are you going to attack?

C: Yes

DECREASES SOLIDARITY: captain committed to perform undesired act 'planned-attack'

Figure 7. Example from [Traum et al. 2005].

Child: what is your favourite fairytale?

HCA: I like several of my fairytales very much. One of my favourites is the ugly duckling. The ugly duckling is not accepted by the other ducks and animals because it looks different. It has a hard time of suffering and troubles. Eventually it wishes only to die till it discovers that it is in fact a beautiful swan. The point is that if you are different then you will get bullied.

Child: can you dance?

HCA: Sorry

Child: can you do the moonwalk?

HCA: I am sorry but i still don't understand. Maybe you are talking about one of the many things i still don't know about

Child: what games do you like to play?

HCA: My favourite childhood game was to play theatre…

Figure 8. Example from [Bernsen and Dybkjær 2005].

Good instruction draws on many information sources, including awareness of students' cognitive and emotional states. Still in its infancy, multimodal emotion recognition and generation is a popular research area today. In the speech modality, prosody delivers cues to the speaker's emotional, cognitive and volitional states, potentially informing the tutor that the student is uncertain, lost, frustrated or saddened by repeated failure, cf. the hand-annotations in Figure 6. Vocal expression of emotional state has been investigated for decades [Scherer 2003]. Among the problems in recognising emotions from speech is that emotions rarely come in a pure full-blown form [Batliner et al. 2003] but must be recognised through

cues expressed not only prosodically but also linguistically. These cues may contribute to detecting trouble in human-computer dialogue [Batliner et al. 2003]. Also in instructional systems, emotion detection is considered important and is actively being researched. For instance, studies indicate that student emotions of frustration and anger correlate with system performance, in particular speech recognition problems [Rotaru and Litman 2006], and ongoing work addresses how student emotion can be automatically detected and used in tutoring dialogue [Ai et al. 2006].

Cognitive modelling is another source of good instruction which manages to select the striking example, or the successful analogy, based on detailed understanding of the student's background knowledge and interests. Training of the social skill of seeing things from the other person's perspective and acting accordingly during negotiation is illustrated in Figure 7 which shows part of a dialogue-turning-bad between a military officer and a virtual doctor. The doctor includes substantial cognitive modelling based on negotiation theory. Regarding the cognitive aspects of student certainness and correctness which are particularly important to instructional systems, studies show that tutors respond differently to student certainty and uncertainty, respectively [Liscombe et al. 2005], and that there is a correlation between uncertainness/incorrectness and recognition problems [Rotaru and Litman 2006], which indicates the importance of asking questions at the right level of difficulty.

The factors mentioned are just some of those that will continue to challenge developers of spoken multimodal instructional systems and their components for a long time to come.

8.2 Spoken Dialogue for Meta-Communication

Our third role for spoken dialogue is for meta-communication, or communication about the communication (interaction) itself, which may be required throughout an instructional session. Under meta-communication we include repetition, correction, clarification, help dialogue and everything to do with introducing the system, its instructional purpose and use. Some types of meta-communication are hard to cope with in today's spoken dialogue systems and better error recovery strategies are very much in demand [Bohus and Rudnicky 2007].

User-requested repetition signals failure to hear or understand what the system said and can usually be handled with a vocabulary that covers the ways in which users might phrase the request. However, understanding failure cannot always be remedied by verbatim repetition. This problem is better avoided through careful output design than resolved on-line. *System-requested repetition* is easy to do and may work if the request is due to simple recognition failure of words and concepts known to the system. However, verbatim user repetition will not work if the input is out-of-vocabulary, grammar or domain. It remains hard for systems to make these distinctions, which is probably why, as remarked by [McTear 2007], much spoken dialogue miscommunication research has focused on speech recognition rather than other error sources. For these other sources, more active strategies are needed, such as asking the user to re-phrase, asking a new question (first system turn, Figure 6) or, like in Figure 8, stepwise nudging the user to change topic or relinquish initiative.

In general, it is worse for the system to misunderstand the user than to fail to understand and ask for repetition or re-phrasing. The former often makes the system appear silly and the user may initiate correction dialogue which can be difficult to handle. Linguistically and conceptually, *correction input* is more diverse than repetition requests, and the system may have to relate the input to what was said several turns back. Moreover, as systems aspire to interpret new types of input information, such as student emotion, new sources of misunderstanding must be dealt with. *System-initiated correction* is ubiquitous in instructional discourse, cf. Paco's turns 4 and 7 (Figure 3), and its non-meta-communication complement, confirmation, is well illustrated in Figures 3 through 6. In fact, constructive and motivating

correction and confirmation design is a major part of instructional systems development. Figure 4 illustrates careful generalised corrective feedback design. Arguably, the main problem is to flexibly handle student input which is not quite right and not quite wrong either.

Clarification is typically a difficult kind of meta-communication dialogue. Clarification requests may require explanation of virtually anything mentioned during dialogue. The best strategy is to try to prevent user clarification requests by design, i.e., by sticking to core-task, core-domain terminology and explaining everything necessary before the student asks. However, this is easier said than done even when the system is being designed for students having well-defined prior knowledge and skills. Everybody can forget the meaning of some technical term but the system can easily make a nuisance of itself by explaining all technical terms as it goes along. Unless the user's potential clarification needs are obvious, this problem has no easy solution and becomes harder the less task-oriented the system is, the wider the domains it covers, and the less is known about the student population. Somehow, future systems must be aware of their own ignorance as illustrated in Figure 8. System requests for clarification are part-and-parcel of instructional discourse but remain hard to do. It is symptomatic that there are no examples in Figures 3 through 8.

Student *requests for help* may concern how to solve a task or operate a device or the system itself. General context-*independent* help is fairly easy to design and may be compared to what we find in GUI help menus. Context-*dependent* help is often more difficult because the task- or discourse context must be taken into account. In Figure 3, having been corrected, the student requests context-dependent help on how to continue from the present state. Since the task context is well-defined, all the system has to do is inform about the next correct action in the context. While help dialogue is generally useful, we are more hesitant recommending spoken dialogue for introducing the system, its instructional purpose and use. Speech is sub-optimal for lengthy and complex explanation and, since students typically use the system for a while, an electronic manual is often preferable.

8.3 Spoken Interaction and Learning Gain

Instruction is all about learning gain. One-to-one human tutoring seems to be very effective compared to classroom sessions. Although sophisticated instructional systems do not achieve the same learning gain as good human tutors, they seem to do better than classroom teaching [Graesser et al. 2005].

Regarding the speech modality, it has been explored if speech recognition problems affect learning gain. Empirical studies have not been able to show negative effects on learning [Pon-Barry et al. 2004, Litman and Forbes-Riley 2005], although recognition problems may cause frustration and affect perceived usability and motivation to use the system.

The impact on learning gain of spoken output quality, including pre-recorded human speech versus synthetic speech, has been investigated by Forbes-Riley et al. [2006] in the context of ITSPOKE (Section 3.4). Learning gain was not influenced by voice quality but this may be due to the fact that the spoken text was also displayed on-screen. This question seems to require more investigation. However, as synthetic voices improve, any negative effects are likely to disappear anyway.

Also the comparative question of learning gain with spoken versus typed text interaction remains an open one [Pon-Barry et al. 2004]. A study by Litman et al. [2006] suggests higher learning gain for spoken human-human tutoring compared to written interaction whereas results for human-computer tutoring are less clear.

Learning gain is probably also strongly related to teaching strategy. Aiming at instructional systems, various studies of human tutoring address what makes a good teacher and which factors influence learning gain. Nevertheless, it remains an open question whether

instructional systems should behave in the same ways as human teachers do. Thus, du Boulay and Luckin [2001] review comparisons of human and computer tutors and of teaching strategies, such as the Socratic approach, and including, e.g., how to deal with errors and how to provide feedback. Some examples of what has been investigated are: Jackson et al. [2004] examined the relationship between dialogue moves and student learning using AutoTutor (Section 3.4). In line with previous research they found that students who received more pumps and hints and played the active part in knowledge construction learned more than those who received more prompts and assertions from a tutor who controlled knowledge construction. Core et al. [2003] looked at initiative using two different teaching strategies and, somewhat surprisingly, found that there is no direct relationship between initiative and learning.

9 Conclusion

Commercial spoken multimodal instructional systems are still rather few, and systems using spoken dialogue in some role or other are fewer still. This no doubt reflects the more general fact that *spontaneous speech* dialogue systems have entered the market only recently, where they are being used to help solve limited tasks of various (non-instructional) kinds. Arguably, most instructional systems which include spoken dialogue will have to handle spontaneous spoken input because it is unrealistic to demand that students learn and remember lengthy sets of fixed keywords and phrases whilst engaged in learning or training things that are difficult enough in themselves. However, with spontaneous speech dialogue systems having entered the market, the technology would seem likely to spread across a wide range of application areas, including instructional systems.

The research systems we have seen provide a useful indication of how far we are. With the technologies illustrated in Figures 3 through 8, it is possible, today, to build useful spontaneous speech, mixed-initiative, multimodal or unimodal teaching and training systems for many different purposes. Today's research systems are typically sufficiently rich in content to allow realistic training or teaching but rarely have the robustness required of commercial systems.

However, several factors seem likely to slow down the proliferation of spontaneous spoken multimodal instructional systems in the near future.

One such factor is speech recognition technology. Recognisers sometimes misrecognise and, although this may not influence learning gain, it is known to cause frustration. Thus companies might be cautious launching applications which include spontaneous speech dialogue. The first computer games with spoken input have not received unanimous acclaim partly because of recognition errors, and the car industry keeps launching spoken keyword-based navigation and other systems rather than spontaneous speech technology. Another factor which we elaborated in Section 4.2, is that adding spontaneous spoken dialogue to instructional systems has relatively high entry costs for researchers and industrial developers alike. The family of technologies required include several components which are poorly standardised as well as being rich in application-specific contents, both of which factors contribute to making development expensive and risky. Thirdly, the market is not necessarily willing to pay a lot for the great opportunity to improve everyone's skills and knowledge through self-training and self-teaching supported by the kinds of spoken dialogue which are ubiquitous in human instruction.

10 Future Research Directions

We have argued that the technologies already exist for including spoken dialogue in a wide range of multimodal instructional systems. To see why the research challenges ahead remain massive, we point out five main reasons why there is still way to go before system instructors can replace good human instructors.

1. Most current spoken dialogue instructional systems address knowledge and skills for which it is relatively easy to determine if the student's problem solving process and/or result is correct. The demands on the system's spoken dialogue capabilities are likely to increase strongly with more open-ended systems where there is no single correct answer and discussion and argumentation is key.
2. Large-domain "real" conversation, as opposed to more or less tightly constrained and primarily system-directed, task-oriented spoken dialogue, is challenged but not conquered by the Andersen system (Figure 8). Human-human conversation follows a multitude of often subtle and sub-consciously practiced principles many of which still have to be demonstrated in running applications.
3. The negotiation trainer (Figure 7) illustrates the tip of another iceberg, i.e., that of emulating human communicative cognition, emotion and volition as it works as an integrated whole, often sub-consciously, during dialogue and conversation.
4. To a larger or smaller extent, human instructors rely on *vision* for watching student task performance, facial expression, gaze, etc. The problems still facing machine vision research imply that most instructional systems must manage without emulating most aspects of human vision for some time to come.
5. Even if it may appear simple to enable natural multimodal student input using, e.g., speech and 2D (surface) or 3D pointing gesture, this is not the case. We are only now discovering the complexity of the multimodal fusion tasks that humans effortlessly accomplish [Martin et al. 2006].

11 References

Ai, H., Litman, D. J., Forbes-Riley, K., Rotaru, M., Tetreault, J., & Purandare, A. (2006). Using system and user performance features to improve emotion detection in spoken tutoring dialogs. *Proceedings of Interspeech ICSLP* (pp. 1682-1685), Pittsburgh, PA, USA.

Allen, J. F. (1979). *A plan-based approach to speech act recognition* (Tech. Rep. No. 131). Toronto, Canada: University of Toronto, Computer Science.

Batliner, A., Fischer, K., Huber, R., Spilker, J., & Nöth, E. (2003). How to find trouble in communication. *Speech Communication*, 40, 117–143.

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

Bernsen, N. O. (2002). Multimodality in language and speech systems - from theory to design support tool. In Granström, B., House, D., & Karlsson, I. (Eds.), *Multimodality in language and speech systems* (pp. 93-148). Kluwer.

Bernsen, N. O., Charfuelàn, M., Corradini, A., Dybkjær, L., Hansen, T., Kiilerich, S., Kolodnytsky, M., Kupkin, D., & Mehta, M. (2004). Conversational H. C. Andersen. First prototype description. In E. André, L. Dybkjær, W. Minker, & P. Heisterkamp (Eds.), *Proceedings of the Tutorial and Research Workshop on Affective Dialogue Systems* (pp. 305-308). Springer, LNAI Series, 3068.

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

Bernsen, N. O., Dybkjær, H. & Dybkjær, L. (1998). Designing interactive speech systems. From first ideas to user testing. Springer.

Bernsen, N. O. & Dybkjær, L. (1999). A theory of speech in multimodal systems. *Proceedings of the ESCA Tutorial and Research Workshop on Interactive Dialogue in Multimodal Systems* (pp. 105-108). Irsee, Germany.

Bernsen, N. O. & Dybkjær, L. (2001). Combining multi-party speech and text exchanges over the internet. *Proceedings of the 7th European Conference on Speech Communication and Technology (Eurospeech)* (pp. 1189-1192). Bonn, Germany: ISCA.

Bernsen, N. O. & Dybkjær, L. (2005). Meet Hans Christian Andersen. *Proceedings of Sixth SIGdial Workshop on Discourse and Dialogue* (pp. 237-241). Lisbon, Portugal.

Bernsen, N. O. & Dybkjær, L. (in press). *Multimodal Usability.*

Bernsen, N. O., Hansen, T., Kiilerich, S. & Madsen, T. (2006). Field evaluation of a single-word pronunciation training system. *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC)* (pp. 2068-2073). Genova, Italy.

Biswas, G., Schwartz, D., Leelawong, H., & Vye, N. (2005). Learning by teaching. A new agent paradigm for educational software. *Applied Artificial Intelligence, 19*(3-4), 363-392.

Bolt, R. A. (1980). Put-that-there: Voice and gesture at the graphics interface. *Proceedings of the 7th Annual Conference on Computer Graphics and Interactive Techniques* (pp. 262-270). Seattle, USA.

Bohus, D. & Rudnicky, A. (2007). Sorry, I didn't catch that. An investigation of non-understanding errors and recovery strategies. In L. Dybkjær & W. Minker.

Cassell, J., Sullican, J., Prevost, S., & Churchill, E. (Eds.). (2000). *Embodied conversational agents*. Cambridge, USA: MIT Press.

Clark, B., Fry, J., Ginzton, M., Peters, S., Pon-Barry, H. & Thomsen-Gray, Z. (2001). A multimodal intelligent tutoring system for shipboard damage control. *Proceedings of International Workshop on Information Presentation and Multimodal Dialogue* (pp. 121-125). Verona, Italy.

Clark, B., Lemon, O., Gruenstein, A., Bratt, E. O., Fry, J., Peters, S., Pon-Barry, H., Schultz, K., Thomsen-Gray, Z. & Treeratpituk, P. (2005). A general-purpose architecture for intelligent tutoring systems. In J. van Kuppevelt, L. Dybkjær, & N. O. Bernsen (Eds.) (pp. 287-305).

Cole, R., van Vuure, S., Pellom, B., Hacioglu, K., Ma, J., Movellan, J., Schwartz, S., Wade-Stein, D., Ward, W. & Yang, J. (2003). Perceptive animated interfaces: First steps toward a new paradigm for human computer interaction. *IEEE Special Issue on Human Computer Interaction*, 91(9), 1391–1405.

Core, M. G., Moore, J. D., & Zinn, C. (2003). The role of initiative in tutorial dialogue. *Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics* (pp.67-74). Budapest, Hungary.

Delgado, R. L. & Araki, M. (2005). Spoken multilingual and multimodal dialogue systems. Development and assessment. Chichester, UK: John Wiley & Sons, Ltd.

Dybkjær, L. & Minker, W. (Eds.) (2007) *Recent trends in discourse and dialogue*, Springer.

du Boulay, B. & Luckin, R. (2001). Modelling human teaching tactics and strategies for tutoring systems. *International Journal of Artificial Intelligence in Education*, 12(3), 235-256.

Forbes-Riley, K. & Litman, D. (2006). Modelling user satisfaction and student learning in a spoken dialogue tutoring system with generic, tutoring, and user affect parameters. *Proceedings of the Human Language Technology Conference/North American Chapter of ACL* (pp. 264-271). New York, USA.

Forbes-Riley, K., Litman, D. J., Silliman, S., & Tetreault, J. R. (2006). Comparing synthesized versus pre-recorded tutor speech in an intelligent tutoring spoken dialogue system. *Proceedings of the 19th International FLAIRS Conference* (pp. 509-514). Melbourne Beach, Florida, USA.

Graesser, A. C., Lu, S., Jackson, G. T., Mitchell, H., Ventura, M., Olney, A. & Louwerse, M. M. (2004). AutoTutor: A tutor with dialogue in natural language. *Behavioral Research Methods, Instruments, and Computers*, *36*, 180-193.

Graesser, A. C., Person, N., Lu, Z., Jeon, M. G., & McDaniel, B. (2005). Learning while holding a conversation with a computer. In L. PytlikZillig, M. Bodvarsson, & R. Bruning (Eds.), *Technology-based education: Bringing researchers and practitioners together* (pp. 143-167). Greenwich, CT: Information Age Publishing.

Graesser, A. C., VanLehn, K., Rosé, C. P., Jordan, P. W. & Harter, D. (2001). Intelligent tutoring systems with conversational dialogue. *AI Magazine, 22*(4), 39-52.

Grice, P. (1975). Logic and conversation. In P. Cole & J. L. Morgan (Eds.), *Syntax and semantics, Vol. 3: Speech acts* (pp. 41-58). New York: Academic Press.

Grosz, B. (1974). The structure of task-oriented dialogs. *IEEE Symposium on Speech Recognition: Contributed Papers* (pp. 250-253). Carnegie-Mellon University Computer Science Department, Pittsburgh, USA.

Grosz, B. & Sidner, C. (1986). Attentions, intentions, and the structure of discourse. *Computational Linguistics, 12*(3), 175-204.

Hansen, T. K. (2006). Computer assisted pronunciation training: The four 'K's of feedback. *Proceedings of the 4th International Conference on Multimedia and Information and Communication Technologies in Education (m-ICTE)* (pp. 342-346). Seville, Spain.

Hill, R.W., Gratch, J., Marsella, S., Rickel, J., Swartout, W. & Traum, D. (2003). Virtual humans in the mission rehearsal exercise system. *Künstliche Intelligenz, Special Issue on Embodied Conversational Agents, 4*(3), 5-10.

Jackson, G. T., Person, N. K., & Graesser, A. C. (2004). Adaptive tutorial dialogue in AutoTutor. *Proceedings of ITS 2004 Workshop on Dialogue-based Intelligent Tutoring Systems: State of the Art and New Research Directions* (pp. 9-13). Maceio, Brazil.

Lesgold, A., Katz, S., Greenberg, L., Hughes, E., & Eggan, G. (1992a). Extensions of intelligent tutoring paradigms to support collaborative learning. In Dijkstra, S., Krammer, H., & van Merrienboer, J. (Eds.), *Instructional models in computer-based learning environments*, NATO ASI Series F: Computer and System Sciences, 104 (pp. 291–311). Springer.

Lesgold, A., Lajoie, S., Bunzo, M. & Eggan, G. (1992b). SHERLOCK: A coached practice environment for an electronics troubleshooting job. In Larkin, J. H. & Chabay, J. H. (Eds.), *Computer-assisted instruction and intelligent tutoring systems* (pp. 201-238). Erlbaum.

Liscombe, J., Hirschberg, J., & Venditti, J. J. (2005). Detecting certainness in spoken tutorial dialogues. *Proceedings of the 9th European Conference on Speech Communication and Technology (Interspeech/Eurospeech)* (pp. 1837-1840), Lisbon, Portugal.

Litman, D. & Forbes-Riley, K. (2005). Speech recognition performance and learning in spoken dialogue tutoring. *Proceedings of the 9th European Conference on Speech Communication and Technology (Interspeech/Eurospeech)* (pp. 161-164), Lisbon, Portugal.

Litman, D., Rose, C. P., Forbes-Riley, K., VanLehn, K., Bhembe, D., & Silliman, S. (2006). Spoken versus typed human and computer dialogue tutoring. International Journal of Artificial Intelligence in Education, 16, 145-170.

Litman, D. & Silliman, S. (2004). ITSPOKE: An intelligent tutoring spoken dialogue system. *Companion Proceedings of the Human Language Technology Conference* (pp. 5-8). Boston, USA.

Martin, J.C., Buisine, S., Pitel, G., & Bernsen, N. O. (2006). Fusion of children's speech and 2D gestures when conversing with 3D characters. *Signal Processing. Special Issue on Multimodal Human-Computer Interfaces, 86*(12), 3596-3624.

Massaro, D. (2005). The psychology and technology of talking heads: Applications in language learning. In J. van Kuppevelt, L. Dybkjær, & Bernsen, N. O. (Eds.) (pp. 287-305).

McTear, M. (2004). Spoken dialogue technology. Toward the conversational user interface. London, UK: Springer Verlag.

McTear, M. (2007). Miscommunication: Why bother? In L. Dybkjær & W. Minker.

Peters, S., Bratt, E. O., Clark, B., Pon-Barry, H. & Schultz, K. (2004). Intelligent systems for training damage control assistants. *Interservice/Industry Training, Simulation, and Education Conference*. Orlando, USA.

Pon-Barry, H., Clark, B., Bratt, E. O., Schultz, K., & Peters, S. (2004). Evaluating the effectiveness of SCoT: a spoken conversational tutor. *Proceedings of ITS 2004 Workshop on Dialogue-based Intelligent Tutoring Systems: State of the Art and New Research Directions* (pp. 23-32). Maceio, Brazil.

Pressey, S. L. (1927). A machine for automatic teaching of drill material. *School and Society, 25*, 549-52.

Rich, C., Sidner, C. L. & Lesh, N. (2001). COLLAGEN: Applying collaborative discourse theory to human-computer interaction. *AI Magazine, 22*(4), 15-26.

Rickel, J., Lesh, N., Rich, C., Sidner, C. L. & Gertner, A. (2002). Collaborative discourse theory as a foundation for tutorial dialogue. *Proceedings of the Sixth International Conference on Intelligent Tutoring Systems* (pp. 542-551). Springer.

Roberts, B. (2000). Coaching driving skills in a shiphandling trainer. *Proceedings of the AAAI Fall Symposium on Building Dialogue Systems for Tutorial Applications* (p. 150). North Falmouth, USA.

Ruttkay, Z. & Pelachaud, C. (Eds.). (2004). From brows to trust. Evaluating embodied conversational agents. Kluwer.

Rotaru, M. & Litman, D. J. (2006). Dependencies between student state and speech recognition problems in spoken tutoring dialogues. *Proceedings of Joint Conference of the International Committee on Computational Linguistics and the Association for Computational Linguistics (Coling/ACL)* (pp. 193-200), Sydney, Australia.

Scherer K. R. (2003). Vocal communication of emotion: A review of research paradigms. Speech *Communication*, 40, 227-256.

Smith, R. W. (1991). A computational model of expectation-driven mixed-initiative dialogue processing. Unpublished doctoral dissertation, Duke University, USA.

Smith, R. W. & Hipp, D. R. (1994). *Spoken natural language dialog systems: A practical approach*. Oxford University Press.

Sleeman, D. & Brown, J. S (1982). Introduction: Intelligent tutoring systems. In Sleeman, D. & Brown, J. S. (Eds.), *Intelligent tutoring systems* (pp. 1-11). New York: Academic Press.

Sommerville, I. (2006). *Software engineering, 8th edition*. Addison-Wesley.

Stevens, A. & Collins, A. (1977). The goal structure of a Socratic tutor. *Proceedings of the National ACM Conference* (pp. 256 – 263). New York: ACM.

Traum, D., Swartout, W., Gratch, J. & Marsella, S. (2005). Virtual humans for non-team interaction training. *Proceedings of the Autonomous Agents and Multi-Agent Systems Workshop on Creating Bonds with Humanoids*. Utrecht, The Netherlands.

van Kuppevelt, J., Dybkjær, L. & Bernsen, N. O. (Eds.) (2005). *Advances in natural multimodal dialogue systems*. Springer Series: Text, Speech and Language Technology, Vol. 30.

VanLehn, K., Jordan, P., Rosé, C., Bhembe, D., Böttner, M., Gaydos, A., Makatchev, M., Pappuswamy, U., Ringenberg, M., Roque, A., Siler, S., & Srivastava, R. (2002). The architecture of Why2-Atlas: A coach for qualitative physics essay writing. *Proceedings of Intelligent Tutoring Systems Conference*, Springer LNCS Series Vol. 2363 (pp. 158–167).

Weizenbaum, J. (1966). ELIZA - A computer program for the study of natural language communication between man and machine. *Communications of the ACM, 9*, 36-45.

Winograd, T. (1971). Procedures as a representation for data in a computer program for understanding natural language (Tech. Rep. No. 235). Boston: MIT AI.