Dialogue Management

Yorick Wilks, Roberta Catizone, Markku Turunen

COMPANIONS Consortium: State Of The Art Papers 2, January 2006

1. Introduction

This report reviews the state of the art research in Dialogue Management (DM) Systems. Dialogue systems have been around since the 1960's, the best known are conversation programs such as Eliza (Weizenbaum 1966) and Parry (Colby 1973). The approaches we describe are categorised as follows: finite state/dialogue grammars, plan-based and collaborative; however, this division is not perfect, since any system can in the end be implemented as a finite state system, but the distinction can correspond to design approaches versus implementation approaches, since finite state models can be used to implement a variety of approaches independently of the design choice. Again, collaborative models may or may not be plan-based, so this distinction too, is less than firm.

2. Basic Types of Dialogue Management Systems

2.1 Dialogue Grammars and Frames

Dialogue grammars, are systems that identify and represent local or global surface patterns of dialogue or patterns of speech acts (Searle 1969) and their responses. Dialogue grammars, which have a long history (Polany and Scha, 1984; Reichman, 1981; Sinclair and Coulthard, 1975), use prescriptive grammars for pattern sequences in dialogues. The first grammars described the structure of the complete dialogue, from beginning to end, whereas more recent approaches are based on the observation that there are a number of sequencing regularities in dialogues, which are called **adjacency pairs**. It has been proposed that a dialogue is a collection of such pairs (Jefferson, 1972), which describe facts such as that questions are generally followed by answers, proposals by acceptances, denials etc. Digressions and repairs are dealt with by using embedded sequences.

Dialogue grammars are used to parse the structure of a dialogue, just as syntactic grammar rules are used to parse sentences. Phrase-structure grammar rules and various kinds of state machines have been used to implement dialogue grammars. For example the SUNDIAL system, uses a dialogue grammar to engage in dialogue about travel conversations.

Although dialogue grammars have been successfully implemented (Müller and Runger, 1993; Nielsen and Baekgaard, 1992), they have been criticised on the grounds that they lack flexibility both as to deviations in the dialogue as well as portability to other domains.

A significant extension of dialogue grammars are **frame-based approaches**, which have been developed to overcome the lack of flexibility of dialogue grammars. The entities in the application domain are hierarchically modelled, and the system can control the dialogue according to the requirements of those entities. Hulstijn et al. (1996), for example, who developed a theatre booking system, arranged frames hierarchically to reflect the dependence of certain topics (like the details of the performance the user wants to see) on others. In Veldhuijzen van Zanten (1996), a train timetable enquiry system, a frame structure relates the entities in the domain to one another, and this structure captures the meaning of all possible queries the user can make. The point of frames is to try to capture a whole topic of dialogue: Lemon and Peters (Lemon 2001) is essentially a frame system, as is the COMIC DM system (Catizone et al 2003) where it is combined with a specific central system to increase flexibility of response.

2.2 Plan-based and Collaborative Systems

Plan-based approaches take the view that humans communicate to achieve goals, including changes to the mental state of the listener. Utterances are seen not just as strings but as performing speech acts (Searle, 1969) and are used to achieve these goals. The listener has to identify the speaker's underlying plan and respond accordingly. For example, in response to a customer's question of "Where are the steaks you advertised?", a butcher's reply of "How many do you want?" is appropriate, because the butcher understands the customer's underlying plan to buy the steaks (taken from Cohen (1990)). Plan-based theories of communicative action and dialogue (for example: Allen and Perault, 1980; Appelt, 1985; Cohen and Levesque, 1990) claim that the speaker's speech act is part of a plan and that it is the listener's job to identify and respond appropriately to this plan. Plan-based approaches attempt to model this claim and explicitly represent the (global) goals of the task. Plan-based approaches have been criticised on practical and theoretical grounds. For example, the processes of plan-recognition and planning are combinatorically intractable in the worst case, and in some cases they are undecidable. Plan-based approaches also lack a sound theoretical basis. There is often no specification of what the system should do, for example, in terms of the kinds of dialogue phenomena and properties the framework can handle or what the various constructs like plans, goals, etc are. Again, a great deal of conversation and dialogue, as the ATT corpora show, are not about planning or tasks **at all**, they are merely conversation and most of this approach is irrelevant.

Conversational Games Theory (Carletta et al., 1995; Kowtko et al., 1991) uses techniques from both discourse grammars and plan-based approaches by including a goal or plan-oriented level in its structural approach. It can be used to model conversations between a human and a computer in a task-oriented dialogue (Williams, 1996).

A (task-oriented) dialogue consists of a one or more transactions, each transaction representing a subtask. A transaction comprises a number of conversational games, which in turn consist of an opening move, and (sometimes optional) end move. An example is an INSTRUCTION game which consists of three nested games: an EXPLAINING game, a QUERY-YN game, and a CHECKING game. The CHECKING game, for example, can consist of a QUERY-YN and a REPLY-Y or a REPLY-N.

The approach deals with discourse phenomena such as side sequences, clarifications etc. by allowing games to be have another game embedded with in it - a technique which allows for the modelling of the complexity of natural dialogue. This approach also makes clear that there is no firm distinction between these and frame systems of section 2.1, since plans can be represented as frames since the days of Schank's Planning scripts (Schank 1977).

A variant called collaborative approaches is based on viewing dialogues as a collaborative process. Both partners work together to achieve a mutual understanding of the dialogue. The motivations that this joint activity places on both partners motivates discourse phenomena such as confirmation and clarification - which are also evident in human-to-human conversations, though, of course, all this rhetoric fits equally well into a planning view. Collaborative approaches try to capture the motivations behind a dialogue and the mechanisms of dialogue itself, rather than concentrate on the structure of the task. The beliefs of at least two participants will be explicitly modelled. A proposed goal, which is accepted by the other partner(s), will become part of the shared belief and the partners will work cooperatively to achieve this goal.

In the TRAINS-93 dialogue manager, Traum's (1996) model of conversation agency extended Bratman's et al. (1988) Beliefs Desires Intentions (BDI) agent architecture. In the BDI model, actions in the world affect agent's beliefs and the agent can reason about its beliefs and thus formulate desires and intentions. Beliefs are how the agent perceives the world, desires are how the agent would like the world to be, and intentions are formulated plans of how to achieve these desires. Traum states two major problems with the BDI model. He argues that agent's perceptions not only influence its beliefs but also its desires and intentions. Also, the BDI model does not support more than one agent. Traum thus extended the BDI model by incorporating mutual beliefs, i.e. what both agents belief to be true and also let perceptions influence desires and intentions as well as beliefs.

Viewgen (Wilks and Ballim 1991b) is a representational system for modelling agents and their beliefs and goals as part of a dialogue system. It has two types of structures: those for agents that can have views of other agents and entities, and those for entities that have no points of view of

their own. It is based on a virtual machine that nests these entities to any depth required for analysis by nesting either type of object inside the first type: i.e. agents can have perspectives of entities and other agents. The important notion is that nested beliefs (about beliefs and goals) are created only at need and not prestored in advance as in the Cohen, Allen, Perrault-type systems above that compute over goals and beliefs. Other related approaches include Novick and Hansen (1995), Novick and Ward (1993), Chu-Carroll (1996), who extends Sidner's (1992, 1994), and Beun (1996).

3. DM Architectures

3.1 SmartKom

SmartKom (Alexandersson and Becker 2001) is a multimodal dialogue system that combines, speech, gesture and mimics input and output within an overall DM architecture of a Blackboard type, called here a “pool” architecture. One of the major scientific goals of SmartKom is to design new computational methods for the seamless integration and mutual disambiguation of multimodal input and output on a semantic and pragmatic level.

The SmartKom Architecture

- Interface modules: on the input side: there is an audio module, on the output side the display manager
- Recognizers and synthesizers: on the input side, there is gesture recognition, prosody and speech recognition modules, on the output side speech synthesis and the display manager.
- Semantic processing modules: this group of modules comprises or transforms meaning representations: gesture and speech analysis, media fusion, intention recognition, discourse and domain modelling, action planning, presentation planning and concept-to-speech generation.
- External services: the function modelling module is the interface to external services, e.g. EPG databases, map services, and information extraction from the Web

The discourse module receives hypotheses directly from the intention analysis module. The hypotheses are validated and enriched with (consistent) information from the discourse history. During this process a score is computed which mirrors how well the hypothesis fits the history. Depending on the scores by the analysis modules and the score by the discourse modeller, the intention analysis module picks the "best" hypothesis.

3.2 Trindi

The Trindi project (Larsson 2000) proposes an architecture and toolkit for building dialogue managers based on an *information state* and *dialogue move engine*. The Information state of a dialogue represents the information necessary to distinguish it from other dialogues, representing the cumulative additions from previous actions in the dialogue and motivating future action. It can be seen as an attempt to make a finite state system more plausible as a general architecture for DM, when combined with other components expressing the overall “information-state” of the system.

Trindi offers a platform for the formalization of the notion of information state which allows specific theories of dialogue to be formalized, implemented, tested, compared and iteratively reformulated. Key to this approach is the notion of UPDATE of information state with most updates related to the observations and performance of DIALOGUE MOVES.

The Information State Theory of Dialogue Modelling consists of

- A description of the **informational components** of the theory of dialogue modelling including common context and internal motivating factors(common ground, commitments, beliefs, intentions, etc.)
- **Formal representations** of the above components (e.g. as lists, sets, typed feature structures, records, etc)

- A set of **dialogue moves** that will trigger the update of the information state (also correlated with externally performed actions such as particular natural language utterances).
- A set of **update rules** that govern the updating of the information state given various conditions of the current information state and performed dialogue moves including a set of selection rules.
- An **update strategy** for deciding which rule(s) to select at a given point from the set of applicable ones.

There is an important distinction between information state approaches (Cooper and Larsson 1999) and dialogue state approaches. In a dialogue state approach, a dialogue behaves according to some grammar where the states represent the results of a dialogue move in a previous state and each state has a set of allowable moves. The "information" is implicit in the state itself and the relationship it plays to other states.

Here, the **informational components** are not conceived of as monolithic nodes in a transition network (as with dialogue state, but rather as consisting of several interacting components). One could model the mental state of an agent or take a more structural view and model the performance of actions. The **formal representations** for modelling various aspects of the dialogue structure range from simple abstract data types to more complex informational systems such as logics.

3.3 WITAS

This system (Lemon 2001) contains a dialogue interface for multi-modal conversations to the WITAS robot helicopter. The requirements of this dialogue system are:

- Asynchronous
- Mixed-Initiative
- Open-ended
- Involves a dynamic environment

The Dialogue Manager creates and updates an Information State corresponding to a notion of dialogue context. Dialogue moves have the effect of updating information states and moves can be initiated by both the operator and the robot. This system can be seen as a dialogue state/information state hybrid that began with a stack structure like COMIC (see below). They dropped this arguing it was too restrictive because navigation back and forth between different sub-dialogues and topics was impossible because information was lost when stack was popped. To compensate for this, they implemented Version II of the dialogue management system which uses a tree structure of dialogue states (*dialogue move tree*), where edges are dialogue moves and branches represent conversational threads. They also wanted to enrich their domain knowledge and inference methods so they implemented a dynamic hierarchical *task tree.* The *task-tree* grows as part of the developing dialogue context and represents tasks and sub-tasks described by the operator and their temporal ordering. This structure allows for reordering and reference to tasks. They also implemented an inference-based model of the robot's changing abilities, depending on dynamic information about the world and the robot's internal state and location.

3.4 CONVERSE

CONVERSE (Levy et al. 1997) was a machine dialogue system funded by Intelligent Research of London which won the Loebner prize in 1997 (and did not enter again against those it had beaten before in that year). That year was the first in which there was no restriction on the topic of discussion with judges, and CONVERSE covered about 80 topics, which were appropriate to its persona as a young female New York-based journalist. It embodied substantial resources, such as WordNet, the proper names of Collins dictionary etc. It could store the Personal information it elicited from a user and build it into the conversation later. Its topic structures were complex ATN scripts that could be left and reentered appropriately and could generate responses using stored/elicited material.

It had no conventional analysis/DAM/generation division, though it used a commercial statistically based parser to pass input to the ATN's. Its control structure was simple blackboard system in

which the ATN's competed to take control of the generation; these decisions were made numerically based on weights assigned by the closeness of fit of the input to the expected input etc. The system had only limited recovery mechanisms if it was not able to find a topic relevant to the input, and relied of seizing control of the conversational initiative as much as it could. Since this system models only plausible conversation, the dialogue had no application goals of any kind.

3.5 COMIC

COMIC was a Framework Five funded IST project (ended in 2004) which applied research in human-human interaction to human-computer interaction. The application of COMIC was bathroom design and it contained speech and gesture input/output with the use of an avatar to generate facial emotion. The DM in COMIC was designed at the University of Sheffield as a general-purpose dialogue management system, designed so that the domain data is separate from the DAM control mechanism. The domain data is expressed using Dialogue Action Forms (DAFs) which are augmented transition networks – a series of nodes and their connected arcs containing tests and the corresponding actions. In order to create and modify the DAFs, a GUI editor (DAF editor) was developed. With the DAF editor, it is a straight forward process to create and modify DAFs. This allows for a relatively self-contained way to maintain the domain data in a dialogue management system. This method of separating domain data along with the visual aid of editing using a graphical representation is novel. COMIC's information sturctures were modelled on those of the higher functionality CONVERSE system, but not with a blackboard architecture, but the flexibility of a stack system to allow reaccess to "pushed" structures, arguing that the losses experienced with this method in the WITAS system (above) were fact acceptable.
The general purpose nature of the DAM means that it could easily be accommodated to other Dialogue systems with a minimum of application specific reorganisation.

The most important features of the DM in COMIC are:

- It is general purpose;
- It can be re-used in other applications with minimal changes/effort;
- It is able to handle different types of Dialogue Management such as user driven, system driven and mixed initiative dialogues;
- It is able to handle different Dialogue Styles;
- It can deal with topic shift nd topic recovery;
- It includes multi-leveled error handling;

3.6 Agent-based dialogue management

A great deal of work has been done in the field of dialogue management to achieve flexible and robust interaction with compact software agents, and this can be seen as an extension of distributed DM architectures such as Communicator (http://communicator.sourceforge.net/index.shtml) in the US These approaches include the agenda-based dialogue management architecture (Rudnicky et al., 1999) and its RavenClaw extension (Bohus & Rudnicky, 2003), Queen's Communicator (O'Neill et al., 2003), SesaME (Pakucs, 2003) and Jaspis (Turunen & Hakulinen, 2000; Turunen et al., 2005a). In these approaches, dialogue management is often implemented using the object-oriented approach. Most importantly, inheritance is used to separate generic dialogue management from domain specific actions.

The modular agent-based approach to dialogue management makes it possible to combine the benefits of different dialogue control models, such as state-based dialogue control and frame-based dialogue control. Similarly, the benefits of alternative dialogue management strategies, such as the system-initiative approach and the mixed-initiative approach (Walker et al., 1998), can be used together in an adaptive way. Using multiple agents for the same purpose makes it possible to combine rule-based and machine learning approaches (Turunen, 2004).

In the Jaspis architecture dialogue agents are used for various adaptive features. For example, in the AthosMail application (Turunen et al., 2004) dialogue control is performed using two approaches to make the system robust for different users. The first approach uses agents for pragmatic processing and sense annotation, while the second approach utilizes numerous

specialized dialogue agents to make multilingual interaction possible (Salonen et al., 2004). In the timetable domain agent-based dialogue management approach is used to implement features such as truly mixed-initiative dialogues (Turunen et al., 2005b), and multimodal guidance to help novice users to interact with the system and bring system-initiative features to the user-initiative interface (Hakulinen et al., 2005).

In the area of speech-based pervasive computing systems the agent-based approach has been used to implement distributed, concurrent, open-ended, dynamically constructed dialogues that can involve multiple participants (Turunen & Hakulinen, 2003). For example, agents have been used to distribute multimodal dialogues between the server and mobile devices (Salonen et al., 2005; Turunen et al., 2005c), and implement pervasive speech-based and auditory dialogues with technology embedded in the environment (Kainulainen et al., 2005).

There have also been applications in DM of the notion of an autonomous agent based on BDI which was originally introduced as an alternative to full planning that could balance reactive and deliberative behaviour (Bratman, Israel & Pollack'88). BDI has been independently developed as a dialog manager around the world (Ardissono'98, Wallis'01) and claims the advantage that it does intentional behaviour and plan failure in a psychologically plausible manner.

4 DM and ASR language modelling

The present situation in dialogue modeling is in some ways a replay, at a lower level, of the titanic struggle in the early 1990's between linguistic models and the data-driven approach to NLP introduced by Jelinek in MT. The introduction into ASR of so called "language models" –which are usually no more than corpus bi-gram statistics to aid recognition of words by their likely neighbours---has caused some, like Young (2002) to suggest that simple extensions to ASR methods could solve all the problems of language dialogue modeling.

Young describes a complete dialogue system seen as what he calls a Partially Observable Markov process, of which subcomponents can be observed in turn with intermediate variables and named:

- Speech understanding
- Semantic decoding
- Dialogue act detection
- Dialogue management and control
- Speech generation

Such titles are close to conventional for an NLP researcher, e.g. when he intends the third module as something that can also recognise what we may call the *function* of an utterance, such as being a command to do something and not a pleasantry. Such terms have been the basis of NLP dialogue pragmatics for some thirty years, and the interesting issue here is whether Young's Partially Observable Markov Decision Process, are a good level at which to describe such phenomena, implying as it does that the classic ASR machine learning methodology can capture the full functionality of a dialogue system, when its internal structures cannot be fully observed, even in the sense that the waves, the phones and written English words can be. The analogy with Jelinek's MT project holds only at its later, revised stage, when it was proposed to take over the classic structures of NLP, but recapitulate them by statistical induction. This is exactly Young's proposal for the classic linguistic structures associated with dialogue parsing and control with the additional assumption, not made earlier by Jelinek, that such modular structures can be learned even when there are no distinctive and observable input-output pairs for the module that would count as data by any classic definition, since they cannot be word strings but symbolic formalisms like those that classic dialogue managers manipulate. Young assumes roughly the same intermediate objects as linguists but in very simplified forms. So, for example, he suggests methods for learning to attach Dialogue Acts to utterances but by methods that make no reference to linguistic methods for this (since Samuel et al. 19w98) and, paradoxically, Young's equations do not make the Dialogue Acts depend on the words in the utterance, as all linguistic methods do. His overall aim is to obtain training data for all of them so the whole process becomes a single throughput Markov model, and Young concedes this model may only be for simple domains, such as, in his example, a pizza ordering system.

All parties in this dispute, if it is one, concede the key role of machine learning, and all are equally aware that structures and formalisms designed at one level can ultimately be represented in virtual machines of less power but more efficiency. In that sense, the primal dispute between Chomsky and Skinner about the nature of the human language machine was quite pointless, since Chomsky's transformational grammars could be represented, in any concrete and finite case, such as a human being, as a finite state machine.

All that being so, researchers have firm predelictions as to the kinds of DM design within which they believe functions and capacities can best be represented, and, in the present case, it is hard to see how the natural clusterings of states that form a topic can be represented in finite state systems, let alone the human ability to return in conversation to a previously suspended topic, all matters that can be represented and processed naturally in well understood virtual machines above the level of finite state matrices.

There is no suggestion that a proper or adequate discussion of Young's views has been given here, only a plea that machine learning must be possible over more linguistically adequate structures than finite state matrices if we are to be able to represent, in a perspicuous manner, the sorts of belief, intention and control structures that complex dialogue modeling will need; it cannot be enough to always limit ourselves to the simples applications on the grounds, as Young puts it, that the typical system S will typically be intractably large and must be approximated.

The future of DM will in part be a reaction to this territorial dispute between ASR and NLP paradigms, but all will agree that the issues remain 1) the extent to which DM data can be learned, and by more sophisticated methods than the reward structures of Walker and Pieraccini (Walker et al., 1998); 2) the ways in which evaluation methods for dialogue systems, and DM in particular, can be evaluated and 3) the extensions to our concept of dialogue that will be needed to deal with distributed dialogues, over time and space, with computers that will come with the spread of small, embedded, "ubiquitous" devices

5 References

Abelson, R and Schank, R. (1977). Script Plans Goals and Understanding. Lawrence Erlbaum Associates.

Alexandersson, J. and Becker, T. "Overlay as the Basic Operation for Discourse Processing in a Multimodal Dialogue System", in Proc. of the 2nd IJCAI Workshop on Knowledge and Reasoning in Practical Dialogue Systems, Seattle, August 2001.

Allen, J.F. and Perault, C.R. "Analyzing Intentions in Dialogues", in Artificial Intelligence, 15(3):143-178, 1980

Appelt, D.E. "Planning English Sentences", Cambridge, University Press, 1985.

Ardissono, L and G. Boella An Agent Architecture for NL dialog modeling in "Artificial Intelligence: Methodology, Systems and Applications", Springer (LNAI 2256) 1998

Beun, R.J. "Speech Act Generation in Cooperative Dialogue", in Luperfoy et al. (1996).

Bohus, D., Rudnicky, A. RavenClaw: Dialog Management Using Hierarchical Task Decomposition and an Expectation Agenda. In Proc. of the Eurospech 2003: 597-600, 2003.

Bratman, M.E., D.J. Israel, and M.E. Pollack, "Plans and Resource-Bounded Practical Reasoning", in Computational Intelligence, 4, 1988

Carletta, J.H. Carletta, A. Isard, J. Kowtko, G. Doherty-Sneddon, and A. Anderson, "The Coding of Dialogue Structure in a Corpus", in Andernach et al. (1995).

Catizone, R., Setzer, A. and Wilks, Y. Multimodal Dialogue Management in the COMIC Project,Workshop on 'Dialogue Systems: interaction, adaptation and styles of management', European Chapter of the Association for Computational Linguistics(EACL), Budapest, Hungary, April 2003

Chu-Carroll, J. "Response Generation in Collaborative Dialogue Interactions", in Luperfoy et al. (1996).

Cohen, P.R. Morgan, J. and Pollack, M.E. (eds), "Intentions in Communication", MIT Press, Cambridge, Massachusetts.

Cohen, P.R. and Levesque, H. "Rational Interaction as the Basis for Communication", in Cohen et al. (1990).

Cooper, R. Larsson, S., Matheson, C., Poesio, M. and Traum, D. "Coding in Structural Dialogue for Information States". Deliverable D1.1. Trindi Project, 1999.

Colby, K. "Simulations of Belief Systems", In Schank and Colby (1973).

Huilstijn, J. Streetskamp, R. ter Doest, H., van de Burgt, S. and Nijholt, A. "Topics in SCHISMA Dialogues", in Luperfoy et al. (1996).

Hakulinen, J., Turunen, M., Salonen, E.-P. Software Tutors for Dialogue Systems. In Proc. of Text, Speech and Dialogue (TSD 2005): 412-419, 2005.

Jefferson, G. "Side Sequences", in Sudnow (1972).

Kainulainen, A., Turunen, M., Hakulinen, J., Salonen, E.-P., Prusi, P., Helin, L. A Speech-based and Auditory Ubiquitous Office Environment. Proc. of 10th International Conference on Speech and Computer (SPECOM 2005): 231-234, 2005.

Kowtko, J.C. Isard, S.D. and Doherty, G.M. "Conversational Games within Dialogue", in Proc. of the ESPRIT Workshop on Discourse Coherence, University of Edinburgh, 1991.

Larsson, S. and Traum, D. “Information State and Dialogue Management in the TRINDI Dialogue Move Engine Toolkit”, in NLE Special Issue on Best Practice in Spoken Language Dialogue Systems Engineering, 2000.

Lemon, O., Bracy, A. Gruenstein, A. and Peters, S. “The Witas Multi-Modal Dialogue System I”, in Proc. of Eurospeech2001, Aalborg (Denmark), 2001.

Levin, E., Narayanan, S., Pieraccini, R., Biatov, K., Bocchieri, E., Di Fabbrizio, G., Eckert, W., Lee, S., Rahim, M., Ruscitti, P. and Walker, M., ``The AT&T Darpa communicator mixed initiative spoken dialog system," ICSLP 2000, Beijing, China, 16-20 Oct. 2000.

Levy, D., Catizone, R., Battacharia, B., Krotov, A. and Wilks, Y. "CONVERSE: A Conversational Companion", in Proc. of the 1st International Workshop on Human-Computer Conversation, Bellagio, Italy, 1997.

Müller, C. and Runger, F. "Dialogue Design Principles - Key for Usability of Voice Processing", in Proc. of the 3rd European Conference on Speech, Communication, and Technology (EUROSPEECH93), Berlin, Germany, 1993.

Nielsen A. and Baekgaard, A. "Experience with Dialogue Description Formalism for Realistic Applications", in Proc. of the International Conference on Spoken Language Processing (ICSLP 92), Banff, Canada, 1992.

Novick, D.G. and Hansen, B. "Mutuality Strategies for Reference in Task-Oriented Dialogue", in Andernach et al. (1995).

Novick, D.G. and Ward, K. "Mutual Beliefs of Multiple Conversants: A Computational Model of Collaboration in Air Traffic Control", in Proc. of AAAI'93, 1993.

O’Neill, I., Hanna, P., Liu, X., McTear, M. The Queen’s Communicator: An Object-Oriented Dialogue Manager. In Proc. of the Eurospech 2003: 593-596, 2003.

Pakucs, B. Towards Dynamic Multi-Domain Dialogue Processing. In Proc. of Eurospeech 2003: 741-744, 2003.

Polany, R. and Scha, R. "A Syntactic Approach to Discourse Semantics", in Proc. of 10th International Conference on Computational Linguistics, Stanford Uni, California, ACL, 1984.

Reichman, R. "Plain-Speaking: a Theory and Grammar of Spontaneous Discourse", PhD thesis, Department of Computer Science, Harvard University, Cambridge, Massachusetts, 1981.

Rudnicky, A. Thayer, E., Constantinides, P., Tchou, C., Shern, R., Lenzo, K., Xu, W., Oh, A. Creating Natural Dialogs in the Carnegie Mellon University Communicator System. In Proc. of Eurospeech 1999: 1531-1534, 1999.

Salonen, E.-P., Turunen, M., Hakulinen, J., Helin, L., Prusi, P., Kainulainen, A. Distributed Dialogue Management for Smart Terminal Devices. In Proc. of Interspeech 2005: 849-852, 2005.

Salonen, E.-P., Hartikainen, M., Turunen, M., Hakulinen, J., Funk, A. Flexible Dialogue Management Using Distributed and Dynamic Dialogue Control. In Proc. of ICSLP 2004: 197-200, 2004.

Samuel, K., Carberry, S., Vijay-Shanker, K. Dialogue Act Tagging with Transformation-Based Learning, Proc. of 17th International Conf. on Computational Linguistics (COLING-ACL '98), 1998

Schank, R.C. and Riesbeck, C.K. "Inside Computer Understanding", Hillsdale, NJ: Lawrence Erlbaum

Searle, J.R. "Speech Acts: An Essay in the Philosophy of Language", Cambridge, University Press, 1969.

Sidner, C.L. "Using Discourse to Negotiate in Collaborative Activity: an Artificial Language", in AAAI-92 Workshop "Cooperation Among Heterogeneous Intelligent Systems", 1992.

Sidner, C.L. An Artificial Discourse Language for Collaborative Negotiation. Proc of AAAI 1994.

Sinclair, J.M. and Coulthard, M. "Towards an analysis of disourse: the English used by teachers and pupils", Oxford University Press, 1975.

Traum, D.R. “Conversational Agency: The TRAINS-93 Dialogue Manager”, in Luperfoy et al. 1996.

Turunen, M. Jaspis - A Spoken Dialogue Architecture and its Applications. Ph.D. Dissertation, University of Tampere, Department of Computer Sciences A-2004-2, February 2004.

Turunen, M., Hakulinen, J. Jaspis, J.2 - An Architecture For Supporting Distributed Spoken Dialogues. In Proc. of the Eurospeech 2003: 1913-1916.

Turunen, M., Hakulinen, J. Jaspis, J. - A Framework for Multilingual Adaptive Speech Applications. In Proc. of 6th International Conference of Spoken Language Processing (ICSLP 2000): 719-722, 2000.

Turunen, M., Hakulinen, J., Räihä, K.-K., Salonen, E.-P., Kainulainen, A., Prusi, P. An Architecture and Applications for Accessibility Systems. IBM Systems Journal, Vol 44, (3): 485-504, 2005.

Turunen, M., Hakulinen, J., Salonen, E.-P., Kainulainen, A., Helin, L. Spoken and Multimodal Bus Timetable Systems: Design, Development and Evaluation. Proc. of 10th International Conference on Speech and Computer (SPECOM 2005): 389-392, 2005.

Turunen, M. Salonen, E.-P., Hakulinen, J., Kanner, J., Kainulainen, A. Mobile Architecture for Distributed Multimodal Dialogues. In Proc. of ASIDE 2005, 2005

Turunen M., Salonen, E.-P., Hartikainen, M., Hakulinen, J., Black, W., Ramsay, A., Funk, A., Conroy, A., Thompson, P., Stairmand, M., Jokinen, K., Rissanen, J., Kanto, K., Kerminen, A., Gamback, B., Cheadle, M., Olsson, F. and Sahlgren, M. AthosMail - a Multilingual Adaptive Spoken Dialogue System for E-mail Domain. In Proc. of Workshop on Robust and Adaptive Information Processing for Mobile Speech Interfaces, 2004: 77-86.

Veldhuijzen van Zanten, G. "Pragmatic Interpretation and Dialogue Management in Spoken-Dialogue Systems", in Luperfoy et al. (1996).

Wallis, P., Helen Mitchard, Damian O'Dea and Jyotsna Das, Dialog Modelling for a Conversational Agent in "AI2001: Advances in Artificial Intelligence, 14th Australian Joint Conference on Artificial Intelligence" edited by Markus Stumptner, Dan Corbett and Mike Brooks, Adelaide AU, 2001, Springer (LNAI 2256)

Wahlster, W., Reithinger, N. and Blocher, A. "SmartKom: Multimodal Communication with a Life-Like Character", in Proc. of Eurospeech2001, Aalborg (Denmark), 2001.

Weizenbaum, J. "ELIZA - A Computer Program for the Study of Natural Language Communication between Man and Machine", Communications of the Association for Computing Machinery 9, 1966.

Wilks, Y. and Ballim, A. "Beliefs, Stereotypes and Dynamic Agent Modeling", In "User Modeling and User-Adapted Interaction", Vol.1, No. 1, Kluwer Academic Publishers, Dordrecht, The Netherlands, 1991.

Wilks, Y. and Catizone, R. "Human-Computer Conversation", Encyclopedia of Library and Information Science, Vol. 69, Allan Kent (ed.). New York: Dekker, 2001.

Williams, S. "Dialogue Management in a Mixed-Initiative, Cooperative, Spoken Language System", in Luperfoy et al. (1996).

Walker, M., Fromer, J., Fabbrizio, G., Mestel, C., Hindle, D. What can I say? Evaluating a spoken language interface to Email. In Proc. of ACM CHI 98 Conference on Human Factors in Computing Systems: 582-589, 1998.

Young, S. Talking to machines—statistically speaking, Proc. ICSOS02.

Current topics in speech synthesis
Morena Danieli
COMPANIONS CONSORTIUM: State-of-Art Papers 1: January 2006

At the 2003 Eurospeech conference in Geneva, Gerard Bailly, Nick Campbell, and Bernd Mobius, all of them members of the ISCA Synthesis SIG, presented suggestions for the hot topics in speech synthesis (Bailly et al. 2003). The goal of their paper was to question some of the foundations of speech synthesis in order to promote the imaginative jump required to anticipate future key issues in this discipline and related technology.

Actually, speech synthesis underwent a dramatic change of paradigm just a few years ago (Campbell and Black 1996): as the above quoted authors wrote, "ten years ago, for example, it was considered unthinkable to use raw waveform in any but the most primitive of 'station-announcement'-type synthesis systems. Now, the leading producers are all *going concatenative*." (Bailly et al. 2003). In other words, "among the traditional text-to-speech techniques, concatenative synthesis seems at present to have won the race for natural sounding artificial speech" (Quazza et al. 2004). Based on the concatenation of speech segments directly extracted from the natural voice of a speaker, it embeds acoustic-phonetic knowledge into the acoustic units themselves. It makes it easier to implement new languages, and it is more likely to preserve the natural quality of the original voice. Just a few years ago, diphones were the most widely used acoustic units, basing their success on their high combinational power: a relatively small number of diphones allows the generation of any message in a given language (Zovato and Sandri 2001). However, the number of junctions and the need of heavy modification of the prosodic parameters cause the resulting 'robotic' voice. That is why speech synthesis moved from diphone approach to the present concatenative, corpus-based or unit-selection approaches. In fact, this concordance with respect to the adopted methodology is only apparent, because in this research field there are still several different synthesis methods, and the new needs of the emerging application fields may cause some rethnking about crucial issues such as basic units, as we will see below.

In the above mentioned paper, the SynSIG members ranked five different hot topics: at present, two-years later, we can state that most of them are still important issues. We will briefly discuss, and update, them in the remnants of this section.

1. Evaluation

Evaluation was the first item on any speech synthesis wish list a few years ago. Users need some methods of comparison bewteen text-to-speech systems, and some baseline reference to compare them.

A promising advancement in this area was done in 2004, when the two 1200-utterance single-speaker database from the CMU ARCTIC databases were released, and challenged current groups working on speech synthesis around the world to build their voices from these databases. One year later, in January 2005, the same CMU group released two further databases and a set ot 50 utterances for each of five genres, and asked to the competion participiants to synthesize these utterances (Black and Tokuda 2005). The resulting synthesised utterances were presented to three sets of listeners: volunteers, US English-speaking undergraduate students, and speech experts. This effort has been the first experiment of speech synthesis evaluation on common datasets.

However, the evaluation problems is still difficult in the speech synthesis field as it is in any other area of natural language processing. Since the needs of speech synthesis are so varied, the evaluation frameworks are either so general that do not apply well to any real need, or so specific that it results almost impossible to compare two systems on the basis of their performance on a "realistic" task. The problem is related with the objective requirements of evaluation: listening is a subjective experience, for evaluating it 'objectivly' it would be necessary to refer to a complex set of of standards and conceptual frameworks that it is not available at present.

2. Extension of the synthesizers: multilinguality.

Many recent applications of speech synthesis require text-to-speech systems to be able to read mixed-lingual input. In the last five years a couple of strategies have been proposed to allow voices of concatenate speech synthesis to utter foreign words and sentences with a plausible and suitable pronunciation. In the polyglot approach, sets of multilingual databases are built by recording a polyglot speaker. This strategy guarantees a perfect pronunciation of each input word, but the number of languages that can be correctly read is limited by the number of languages known by the speaker. To overcome this problem a second strategy, where a monolingual acoustic database is maintained, has been developed (Badino et al. 2004a, Badino et al. 2004b). Each input word is transcribed according to its language and the phonemes of the resulting transcription have to be mapped onto the phonemes of the phonological system of the language of the monolingual voice. Although its resulting approximated pronunciation is characterized by strong foreign accents, this approach tends to simulate a speaker who makes the choice of maintaining his native-tongue phonological system when foreign words must be pronounced

3. Emotion / Expression

Research on "emotional" aspects of speech production has received a growing interest during the past few years, but not all authors agree on 'emotion' as the best term for this genre, since it may be more suited to denote the paralinguistic aspects of speech recognition. Some authors prefer the label "expressive speech" (Bulut et al. 2002).

Within this area there are some relevant research themes, such as prosody control. At present, the prosody of many state-of-the-art synthesizers falls short of being able to reproduce the variation quality required for emotional speech. If synthesis needs to express emotional variation and a variety of paralinguistic information, the issue of voice quality control becomes a crucial research area.

4. Multimodal aspects of speech synthesis

Multimodality of speech synthesis is another area of increasing growth. Conversational agents need to combine speech with non-verbal communication for generating intelligible multimodal utterances (for example, see Kopp and Wachsmuth 2004, a research focused on the generation of gesture and speech from XML-based descriptions). The integration of voice output, gesture, and facial expression reflects the fact that speech accompanied by visual information provides a more robust and rich way of communicating, particularly in noise environments, and when young people, or the elderly, are involved. Certainly, the integration of speech output systems with facial expressions requires that the first technology is able to meet severe requirements, such as the synchronization of prosody control with eye movements, and these are not present in traditional voice-only application environments. Some interesting

5. Input

Input to the synthesizers is of vital importance: raw texts cannot specify the appropriate paralinguistic interpretation of semantic content. Annotated input to a synthesizer would allow a finer specification of speaking style and of the intended interpretation of a message. During the last few years, some mark-up descriptions, such as SSXML, the VoiceXML based mark up language developed within the VoiceBrowser group of the W3C (Burnett et al. 2004, McGlashan et al. 2004), were widely adopted by most text-to-speech producers. However, the expressiveness of such mark-up languages is still insufficient for representing the rich set of interpretation labels needed to charcterize expressive speech.

Emerging research and application fields.

The issue of basic units for speech synthesis is an old one (Nebbia et al. 1998), but many researchers postulate that it will once again come to the fore, because the granularity of the unit that is used for selection is an crucial aspect if speech synthesizers are to be driven by voice input (Bailly et al. 2003). It is an open research issue whether the granularity of the unit used for selection is better determined by the spectral or articulatory features, rather than by phoneme-based definitions. An advance in this research area

could provide a dramatic improvement in synthesized voice quality, perhaps the improvement needed to approach the demanding, emerging application fields.

Some new business areas, such as entertainment, customer-care, robots, education, home and car automation, to list just a few of them, may require more than a good voice quality for the speech synthesizer of the future. As the output quality of speech synthesis improves, we can expect that it will be required to replace the (recorded) human voice in many human-computer interactions. In such cases it will be also required to express personality. If the next generation speech synthesizer is to be used in unobtrusive conversational interaction with human interlocutors, there will be a need for expression of moods and attitudes, and more use will be made of 'fillers' such as laugh, cough, filled pauses, etc...(see Hamza et al. 2004 for example, and Zovato et al. 2004 for an alternative approach) The research in this area requires that the speech synthesis technologies can model and embody psychological attitudes (Scherer 2000).

References

L. Badino, C. Barolo, and S. Quazza, (2004 a) "A General Approach to TTS Reading of Mixed-Language Tests", in *Proceedings of ICSLP 2004*, Jeju Island, South Korea, October, 2004.

L. Badino. C. Barolo, and S. Quazza, (2004 b) "Language Independent phoneme mapping for foreign TTS", in: *Proceedings of the 5th ISCA Speech Synthesis Workshop*, 2004, Pittsburg, USA.

G. Bailly, W.N. Campbell, and B. Mobius, (2003) "ISCA Special Session: hot topics in speech synthesis", in: *Proceedings of Eurospeech 2003 – Geneva"*, Geneva, Switzerland, 2003, pp. 37-40.

A.W. Black and K. Tokuda, (2005) "The Blizzard Challenge – 2005: Evaluating corpus-based speech synthesis on common datasets", in: *Proc. of INTERSPEECH 2005*, Lisbon, Portugal, Sept'05, pp. 77-80

Bulut, M., S.S. Narayanan, and A.K. Syrdal (2002), "Expressive speech synthesis using a concatenative synthesizer", in *Proceedings of ICSLP 2002*, Denver, Colorado, October, 2002

D. C. Burnett, M.R. Walker, and A. Hunt (Eds), (2004), *Speech Synthesis Markup Language (SSML) Version 1.0*, W3C Recommendation, 7 September 2004, http://www.w3.org/TR/speech-synthesis/

W.N. Campbell and A.W. Black, (1996) "Prosody and the Selection of Source Units for Concatenative Synthesis", in: J. van Santen et al. (eds.), *Progress in Speech Synthesis*, pp. 279-292, Springer New York, 1996.

S. McGlashan et al. (Eds) (2004) *Voice Extensible Markup Language (VoiceXML) Version 2.0*, W3C Recommendation, 16 March 2004.

W. Hamza et al. (2004), "The IBM expressive speech synthesis system", in *Proceedings of the 5th ISCA Speech Synthesis Workshop.* 14-16 June 2004, Pittsburgh, Pa.

S. Kopp and I. Wachsmuth (2004), "Synthesizing multimodal utterrances for conversational agents", in *Computer Animation and Virtual World*, Vol. 15, Issue 1, pp. 39-52

L. Nebbia, S. Quazza and P.L. Salza, (1998) "A Specialized Speech Synthesis Technique for Application to Automatic Reverse Directory Service", *Proceedings of IVTTA '98*, Torino, September 1998, pp. 223-228.

K.R. Scherer, (2000) "A cross-cultural Investigation of emotion inferences from voice and speech: Implications for speech technology", in: *Proceedings of ICSLP 2000*, Beijing, China, October 2000.

S. Quazza, L. Donetti, L. Moisa, and P. L. Salza, (2004) "ACTOR®: A multilingual unit-selection speech synthesis system", in: *Proceedings of the Fourth ISCA Workshop on Speech Synthesis (SSw4)*, Pitlochry, Scotland, September 2004

E. Zovato, and S. Sandri, (2001) "Two feature to check phonetic transcriptions in Text To Speech Systems", *Proceedings of EUROSPEECH 2001*, Aalborg, Vol. 3, pp. 2243-2246.

E. Zovato, A. Pacchiotti, S. Quazza, and S. Sandri (2004) "Towards emotional speech synthesis: a rule based approach", in *Proceedings of the 5th ISCA Speech Synthesis Workshop.* 14-16 June 2004, Pittsburgh, Pa.

Emotion in Human-Agent Interfaces

Peter Wallis, Roger Moore, Petra Fagerberg, Marc Cavazza and Yorick Wilks

COMPANIONS Consortium: State Of The Art Papers 3, January 2006

Recent years have seen an upsurge of interest in the implications and applications of emotional behaviour in the human-computer interface (Picard 1997). In particular, research has focused on interactive systems involving visual and/or vocal modes of communication, where the user and/or the system may exhibit both positive and negative expressive behaviour in the course of an ongoing dialogue. Such behaviour can lend naturalness to an otherwise artificial situation, but it can also be exploited to enhance the communicative effectiveness of a human-agent interface; for example, a system may detect the emotional state of a user (from the look on their face, or from the tone of their voice) and act accordingly, or it may signal a problem by adopting suitable emotional behaviour itself (by generating appropriate facial and/or vocal expressions).

Current research in this area revolves around virtual manifestations of human-agent interfaces known as 'Embodied Conversational Agents' (ECAs): speech-enabled animated characters that can act for example as a receptionist, a tour guide or a personal tutor (Cassell et al 2000, Pelachaud and Poggi 2001, Beskow et al 2005). The ability of an ECA to detect and exhibit emotional behaviour is believed to be an important requirement for achieving effective and naturalistic human-computer interaction (Bates 1994).

Emotion and Behaviour

The formal study of emotion in human (and animal) behaviour has a long history, from the early observational work of Charles Darwin (1872) up to the recent emergence of 'Affective Science' (Davidson et al 2003) as a cohesive discipline. Over that period, three main categories of psychological model of human emotion have emerged.

The earliest 'discrete' theories of emotion (stemming from Darwin's work) hypothesised the existence of a small number of basic emotions, such as happiness, sadness, fear, anger, surprise and disgust (Ekman 1999). In such theories, it is supposed that these emotions are based on specific physiological response patterns to external stimuli.

Another early model of emotion is the 'dimensional' approach (Wundt 1874) in which a wide variety of emotions are mapped into a low-dimensional space. For example, the coordinates of a hypothesised two-dimensional representation of emotion would reflect subjective aspects of behaviour such as feelings of positivity vs. negativity and active vs. passive. Cowie et al (2001) use such a scheme as the basis for FEELTRACE, a computer-based tool for emotion annotation.

The third, and most recent, theoretical view of emotion is the 'componential' model. This approach emphasises the variability of different affective states, and links the production of an emotion to the appraisal of a situation with respect to an organism's needs and goals (Scherer 2001).

Both dimensional and appraisal-based approaches have been sources of inspiration for affective interface systems. The former have been generally used to augment traditional user interfaces, while the latter have been used to define agent's emotion in terms of their internal cognitive mechanisms (e.g. anticipation of plan success or failure, see (Gratch, 1999)).

Previous Relevant EU-funded initiatives

SAFIRA (IST-1999-11683) was one of the first EU-funded projects specifically dedicated to affective computing. Its objectives were to create a framework to enrich interactions and applications with an affective dimension. It planned to implement a toolkit for affective computing combining a set of components addressing affective knowledge acquisition, representation, reasoning, planning, communication and expression. Another objective of the project was to verify under which conditions the hypothesis that emotion, as well as other affective phenomena, contributes to improve rationality and general intelligent behaviour of the synthetic characters, thus leading to more believable interactions between humans and computers. The project has released a toolkit and has generated a number of demonstrators: "FantasyA", a novel kind of affective computer game, the "Influencing Machine", an affective ambient intelligence system and "James the Butler", an affective ECA. James is a Personal Sales Assistant (PSA) that employs intelligence and adaptive reasoning methods to provide active, collaborative assistance to a customer of an online store. James is an agent that amplifies or modifies the

motivational state of an agent and its perceived bodily state. It has the ability to perceive and produce the visual (animated expressions), verbal and non-verbal signals and regulate the flow of information between service agents, the interface agent and the user. These capabilities enable James to engage in complex interactions with customers via natural social communication rather than complex command languages, or direct manipulations.

NECA (IST-2000-28580) was dedicated to multi-modal communication with animated synthetic personalities. A particular focus in the project lies on communication between animated characters that exhibit credible personality traits and affective behavior. The key challenge of the project is the fruitful combination of different research strands including situation-based generation of natural language and speech, semiotics of non-verbal expression in situated social communication, and the modelling of emotions and personality. The project developed several demonstrators, among which the eShowroom, that featured ECA as car sales assistants. In this demonstrator, personality profiles could be defined for agents that determined inter-agent dialogues.

MAGICSTER (IST-1999-29078) was concerned with the development of believable conversational interface agents which make use of gaze, facial expression, gesture and body posture as well as speech in a synchronised fashion. The project also intends to evaluate the use of conversational agents in laboratory conditions to determine which aspects of the embodied agent are relevant to various situations of human-computer interaction. Finally, the project aimed to develop and document the agent architecture and components to enable other research and development teams to prototype and evaluate new versions of the agent interface in new domains and for novel tasks. MAGICSTER contributed important developments to the field of affective ECA, such as the APML markup language. The project produced several demonstrators of emotional dialogues with ECA (Cavalluzzi et al., 2003) some integrating state-of-the-art ECA (the "Greta" agent) with state-of-the-art dialogue toolkits (the TRINDI system). The emotional model was derived from the OCC approach.

PF-STAR (IST-2001-37599) addressed three crucial areas of user interfaces: technologies for speech-to-speech translation, the detection and expressions of emotional states, and core speech technologies for children. The project has produced a variety of results in the field of affective dialogue; on the relation between users' emotional states and dialogue progression (Batliner et al., 2004) and on how emotions influence articulatory patterns with application to MPEG-4 encoding of talking faces (Beskow, 2004). The system has produced several prototypes and toolkits to develop talking heads supporting emotional display (Cosi et al., 2004).

VICTEC (IST-2001-33310) developed ECA as part of a tutoring system educating children on bullying issues. It is based on the generation of empathy towards virtual actors being bullied by fellow children as part of an interactive narrative. The FearNot! System developed as part of the project includes NL communication features supporting the recognition of speech acts to give advice to the virtual actors. The prototype has been used to stage real-world evaluations in schools.

NICE (IST-2001-35293) was a Human Language Technology project developing Multimodal conversation with virtual characters in an edutainment context. It was dedicated to children and teenagers, introducing them to a virtual character of H.C. Andersen. NICE implemented Multimodal dialogue based on speech and gesture recognition.

HUMAINE (IST-2004-507422). The rapid growth of academic research in affective computing has resulted in the formation of an EU-funded network of excellence: 'Human-Machine Interaction Network on Emotion' (HUMAINE) funded under the Framework 6 IST Programme. Started in 2004, HUMAINE involves around thirty different laboratories and incorporates research across a wide number of disciplines and applications. The project's research is organised into six thematic areas:

- theories of emotion
- from physical signals to emotionally significant features and vice versa
- patterns of signs that convey emotion in interactions
- functions of emotion-related elements in communication and persuasion
- emotion in cognition and action
- usability of emotion-oriented systems

HUMAINE aims to contribute to "the development of systems that can register, model and/or influence human emotional and emotion-related states and processes".
Computers obviously do not experience emotion in the same way as people, but less obviously perhaps,

emotion plays an essential role in the everyday doings of human beings. At first blush, the emotions that come to mind are those Cowie and Schröder (2005) call **episodic emotions** and include fear, happiness, anger and so on. The categories of emotion of primary relevance to our day to day doings are however the **pervasive emotions**. These include moods (cheerful, gloomy, irritable, listless, depressed, buoyant) Interpersonal stances (distant, cold, warm, supportive, contemptuous) attitudes (liking, loving, hating, valuing, desiring) and affect dispositions (nervous, anxious, reckless, morose, hostile). These emotions are pervasive in the sense that the decisions we make every minute are influenced by some underlying mechanism that surfaces in behaviour that is described in these terms.
Cowie and Schröder talk of the watershed issue for HUMAINE being an understanding of 'really natural language processing' and emphasise the fact that a conversation with a machine is an affective loop for which, like playing a drum (to use Höok's example) the user behaviour should be reflected in the user experience. Direct applications of an understanding of emotion in dialog would include things like trouble shooting in automated call handling systems - if the agent could recognise when the caller has (or was about to have) a negative experience, what conversational strategies would relieve or avoid the situation? This issue has been addressed directly (see Brahnam (2005) for example) but is there some underlying explanation for why and when we experience emotion, and why and when we expect others - including synthetic characters and ECA - to experience emotion?

Emotion and Interaction

As outlined above, the HUMAINE umbrella includes many different approaches to studying emotion, some of which are based on a strategy that has proved so successful in computational linguistics over the last decade. By tagging a corpus of speech or text - or in the case of ECA, of video - with markers indicating the emotion being expressed at that point, machine learning techniques or statistical analysis can, it is felt, tell us something important about the nature of emotion in human interaction. Marking up corpora for emotion is a much more dubious task than marking up for part-of-speech, and its consistency and value has not yet been shown. .
However, significant activity has been dedicated to emotion markup langages and other a priori definitions of emotional categories that would be used to control the animation of virtual agents

Mel Slater and colleagues (2000) have designed a system, Acting in Virtual Reality, where the user expresses her emotions by changing the characteristics of a drawn face. The user can influence the eye brows and the mouth of that face. The mouth can express happy, neutral and sad, while the eye brows can express surprised, neutral and angry. By interacting with both the mouth and the eye brows at the same time the user can create more complex expressions. It is also possible for the user to affect some body parts of her avatar. The emotions are carried out by the user's avatar in a virtual rehearsal system. The system was set up to be used by actors to see if they could rehearse a play in virtual reality that in the end was going to be performed on a real stage. The actors who were not previously familiar with each other met in the virtual reality four times for a one hour rehearsal each time. Then they met a fifth time for a live performance in front of an audience. Even though the actors did not think the system could replace real rehearsal they all learnt to master the program and to use its qualities. One of the actors compared it to talking on the telephone which is not like a real life meeting but still very effective and interesting.
ExMS is a system where the user explicitly states her emotions (Persson 2003). ExMS is an avatar-based messaging system where users can create short pieces of animated film to send to each other. The idea is that each user chooses an avatar that she can identify herself with and by using the library of animated expressions specific to her character she can express feelings, reactions and moods in the messages she sends to her friends. One disadvantage with ExMS was that the avatars had so much character in themselves so that it sometimes was hard for people to see themselves and their own expressions represented through their avatars.
Another communicational system where users explicitly state the emotion they want to communicate is used in CHATAKO, a speech synthesis system developed to assist people with communication problems (Iida et al. 2000). In CHATAKO the user writes what she wants to say and then chooses if she wants to say it with a female or a male voice and what emotional value she wants that voice to have. The prototype has three emotions to choose from; joy, anger and sadness.
These two systems are examples of applications that support emotional communication. The first two are creative and fun and CHATAKO is an important solution for people with speech problems. However, they do not fulfill the physicality and ambiguity of an affective loop that we want to create. Even if the expressivity in Acting in Virtual Realty and CHATAKO was to be extended on there were other problems with personality and open interpretation experienced in ExMS where there were more expressions to choose from.

The principal applications making use of emotional ECA consist of Information Access (e.g. user emotion recognition in SMARTKOM, (Streit et al., 2004)), Training and Tutoring systems, as well as entertainment systems (i.e. interactive storytelling featuring ECA as virtual actors).
We can identify from the state-of-the-art a set of important research problems that characterise ECA-based affective interfaces. These problems can be listed as:

- *Empathy* is an important concept to characterize the quality of the relation established between users and ECA. The benefits of Empathy in communication have been reviewed by (Martinovski et al., 2005) who also studied the linguistic manifestations of empathy (or its rejection) in dialogue. It should also be noted that depending on the application, empathy can be exerted by the ECA towards the human (Prendinger et al., 2004) or by the human towards the ECA (as in the FearNot! System introduced above).
- *Politeness* is a recurring topic in ECA. Politeness can be an important part of a information access system or a tutoring system, although in some cases the tutoring system precisely consists in acquiring elements of politeness and proper behaviour, for instance in another language and culture, as in the Tactical Language Training System developed at the University of Southern California. It should be noted that both sides of politeness are explored: agents politeness (Johnson et al., 2004) and users politeness (or rudeness, see (Kopp et al., 2005)).
- *Humour*. The importance of humour for ECA has been discussed as early as 2002 by Nihjolt (2002), who related its implementation to the use of emotional modelling (e.g. using the OCC model). ECA typically involve multimodal humour through co-ordinated non-verbal behaviour and humorous language generation. Humour is however not limited to "jokes"; in some cases humour can also be represented by witticisms and punchlines, for instance within an interactive storytelling context featuring virtual actors (Cavazza et al., 2004) (Cavazza and Charles, 2005).
- *Naturalness*, which can be defined as the endeavour to provide believable communication rather than "system prompts" uttered by an agent, tends to be taken for granted in recently developed systems and to disappear behind more specific relational modes, such as those listed above.

.

Emotion in Language and Speech

According to Scherer (2003), systematic research into the effect of emotion on the voice started in the 1960s when psychiatrists took an interest in diagnosing affective states from vocal expressions. Since then, psychologists, linguists, phoneticians and engineers have also become involved, culminating in 2000 with an international workshop on 'Voice and Emotion' and a subsequent special issue of the journal Speech Communication (Douglas-Cowie et al 2003).
One of the first demonstrations of adding emotion to synthetic speech resulted from the work conducted by Ian Murray at Dundee University (Murray and Arnott 1993). Murray developed a system called 'HAMLET' which used rules to alter voice, pitch and timing of the commercially available DECTalk speech synthesiser. DECTalk is an example of a 'formant-based' text-to-speech system which has the characteristic that it provides parametric control over the content and characteristics of the output voice, hence it is easy to manipulate in order to introduce expressive behaviour. Similar work took place at MIT (Cahn 1990) and other early work combined the generation of both vocal and facial expression (Henton and Litwinowicz 1994). In his review of emotional speech synthesis, Schröder (2001) concluded that it was not yet usable in many real life applications due to the restriction of using a few basic emotions, the simplicity of the models for intonation and timing and the lack of naturalness of formant-based speech synthesisers.
Recent years have seen a significant upsurge of interest in the recognition of emotion in speech, particularly for applications such as call-monitoring for 'Interactive Voice `Response' (IVR) systems, e.g. to detect a complaint. As in speech synthesis, much of the relevant research is directed towards the identification of a few basic emotions such as happiness, anger and sadness. Such investigations fall under the wider heading of the recognition of 'Speech under Stress' (Moore 1996); a broad area of research that encompasses the full range of external and internal 'stressors' that can condition the speech production process (for example, cognitive load, physical and mechanical stressors, physiological constraints, psychological stress and emotion). Interest has also been fuelled by recent coverage in the popular press of how financial institutions such as insurance companies are beginning to use so-called 'Voice Stress Analysis' (VSA) on incoming telephone calls in order to detect (and deter) deception - although it must be stated that scientific opinion is divided about the effectiveness of such devices (Haddad et al 2002).
Many studies on the recognition of emotion in speech have investigated the use of prosodic features, i.e. the influence of emotion on the patterning of the pitch, intonation and timing in speech signals. More

general stress analysis techniques also use spectral features, and much experimental work has been done using a variety of different pattern classifiers such as linear discriminant analysis (LDA), artificial neural networks (ANNs) or Gaussian mixture models (GMMs). In general, such research has shown that, given sufficient training data, it is possible to classify the emotional content of speech with a reasonable degree of accuracy, e.g. ~65% for four basic emotions (Yildirim et al 2004). In a recent study, Rigoll et al (2005) demonstrated an ability to classify utterances taken from an automotive application with an accuracy of 74% using acoustic information only, 60% using linguistic information only, and 83% using both (for seven basic emotions).

Emotional Content and Dialogue Management

Emotion as a computational topic in artificial intelligence goes back to the speculations of Sloman (REF) and the implementation of Colby's PARRY (Colby 74), the first robust conversational agent in the early 1970's which had a simple screen dialogue interface. PARRY had internal variables FEAR and ANGER which rose and fell according to keywords in the human dialogue partner's input, and whose values selected a response at each turn from the alternatives available as responses to the given input. From the late 1980s there was a general consensus in the natural language processing community that assembling the knowledge about Language and the world was going to be a key problem for any useful natural language system. The solution, for those interested in whole systems, was to have their conversational agent play a constrained role. Hence, PARRY played the role of a patent with paranoia, and Weizenbaum's ELIZA (Weizenbaum 1966) program played the role of a Rogerian psychologist. Later work also focused on roles that, at first blush, could be construed to have no emotional content such as a virtual teacher (Zukerman 2001) or virtual assistants in the form of the ever patient library assistant or (un)smiling butler (AskJeeves). As discussed above however, the pervasive emotions influence our moment by moment decisions, and in order to make virtual tutors and advisors believable, these virtual humans must have some understanding of the human emotional landscape. Capturing such information is key to creating believable characters of any kind. Those coming from a background of dialog systems have however have tended to see the issue as one of **politeness**.

Walker et al (97) introduce Linguistic Style Improvisation in which texts are generated by decomposing a conversational goal into a series of speech-acts. Parameters to this planning process are social distance and power relations that must be accounted for in the management of politeness. REA (Cassell 99) is a virtual real-estate agent that shows real humans through virtual houses. In order to do this REA must manage her interpersonal relationship with the human. The role of real-estate agent is more flexible than the formal roles of teacher or Rogerian psychologist, and the changing role is reflected in REA's measure of social distance. Franco (Estival 03) is an ECA designed as a virtual assistant in a military command and control centre. Such an agent must not only be able to convey information, it must do it in a manner that obeys the rules of social relations. Once again the approach taken was to see the problem as one of politeness (Wallis 01). Recent work within the scope of HUMAINE has embraced the notion of politeness and Rehm and André (05) are marking up video with politeness markers.

A more recent trend has been the adoption of more sophisticated models of emotions based on cognitive modelling, such as the OCC model (Ortony et al., 1988) and the Oatley model (Oatley and Johnson-Laird, 1987). This can be explained as a consequence of the development in the past few years of more sophisticated ECA comprising cognitive, affective and communicative modules.

One illustration of this approach is given by research carried out at the Institute of Creative Technologies (ICT) of the University of Southern California, in particular the Mission Rehearsal Exercise (MRE) project. This military training application is dedicated to crisis situations in peacekeeping operation and features emotional ECA throughout the cast of virtual actors, from military personnel reacting to the trainee's orders or shifting blame on fellow virtual agents, to civilians caught in dramatic situations as part of the mission being rehearsed. The MRE system features emotional ECA that incorporate user emotion detection, ECA emotional modelling related to the agent's cognitive model and a complete spoken dialogue system. The system has been used to experiment the role of emotion in dialogue (Traum et al., 2004). In addition, its various components are sufficiently integrated to support the processing of global communication situations such as social judgements and attribution of responsibility (Mao and Gratch, 2004), modelling how agents cope with situations (Marsella and Gratch, 2003), etc.

Several projects have beed dedicated to emotional ECA at the Center for Advanced Research in Technology for Education (CARTE) of the University of Southern California, with specific emphasis on ECA with social skills, politeness, etc. For instance, the tactical language training system (Johnson et al.,

2005) develops tools to support individualized language learning, and apply them to the acquisition of tactical languages: subsets of linguistic, gestural, and cultural knowledge and skills necessary to accomplish specific tasks. Other projects have been dedicated to the impact of emotions in learners using Interactive Tutoring Systems featuring ECA.

The "Façade" system (Mateas and Stern, 2004) is an interactive narrative application, in which the user communicates through Multimodal dialogue with artificial actors evolving in a virtual 3D world. The main mode of user interaction is through (written) dialogue with the virtual characters; the multimodal aspects corresponding to the physical interactions resulting from user navigation in the virtual world (which select deictic references to virtual world objects or characters). The user input is interpreted as speech acts that in turn affect the emotional status of the virtual actors, altering the course of the narrative.

Other recent research has developed emotional models to assist the control of user-system dialogue itself, and this shall be of particular relevance to future ECA systems, especially those based on spoken dialogue systems.

- Bosma and Andre (2004) have introduced a method to disambiguate dialogue acts using emotional data, with the goal of improving the communication between users and ECA. The underlying hypothesis is that the meaning of utterances in dialogue is closely correlated to the user's emotional status, and that the emotional status can even assist in determining the actual meaning of two textually identical utterances. Their system processes natural language input through finite-state systems that contain variable weights for various dialogue acts and obtains these weights from a Bayesian network computing the user's emotional status (using application history as well as input from physiological sensors).
- Riccardi and Hakkani-Tur (2005) have studied the grounding of emotions in spoken dialogue systems in the context of AT&T's "How May I Help You System" (Gorin et al., 2002). This work essentially addresses the impact of the user state on machine dialogue strategies and machine's performance. Although the emotional model is restricted to a positive/negative rating (towards the system), this work demonstrated that the recognition of emotional patterns (including temporal patterns) could be used to improve dialogue strategies and to predict speech recognition errors.

Developing dialog systems for ECA that more than simply demonstrate the possibility takes considerable resources. For instance, one long running project at the University of Bielefeld that has developed Max (Kopp et al, 05). A recent incarnation of Max has been running for 18 months as a guide in the Heinz Nixdorf Museums Forum in Paderborn, Germany. Max incorporates an emotion system which keeps track of, for example, obscene or politically incorrect wordings in the user input. Repeated insults will put the agent in an extremely bad mood which eventually results in Max leaving the scene. This abusive behaviour is not bought on by a lack of politeness on behalf of the agent; it appears to be spontaneous and extremely common amongst dialog systems in public places. These recent findings close the loop between the emphasis on politeness in the earliest dialogue systems (PERRY, ELIZA) It seems something is missing from our common understanding of what natural language processing is about.

The Companions approach to Emotion

The paradigm we advocate is somewhat different from the popular corpus and mark-up approach, and rests on investigating and modelling not human-human behaviour involving emotions but on the behaviour of humans to machines, which is now an established and attested type of behaviour, and what behaviour in humans towards machines we should be seeking to elicit in a cooperative interface like a COMPANION (Wilks 2004). This interaction type is quite distinct (de Angeli 01) and human-human dialogue corpora may not be relevant to its modelling.

The intention is to embrace the notion of an 'affective loop' and develop a series of spoken language dialog systems that provide varying degrees and kinds of (pervasive) emotional feedback. Rather than "understanding" what is being said, or grounding information state updates (for example, Kreutel 2000) the primary aim will be to provide feedback that engages the user in the process of conversation. In other words, the COMPANIONS approach aims at grounding emotional aspects in the semantic content of human-agent communication, rather than resorting to an a priori ontology of affects.

From a technical perspective, a further challenge consists in developing a proper HLT approach supporting this process of engagement through dialogue. This approach should include semantic principles underlying the automatic processing of various aspects of human-ECA bonding, including

politeness, empathy, humour, etc. These semantic principles should also underlie the representation of dialogue acts in a way which would be compatible with the experiments in dialogue control planned in COMPANIONS.

Like playing a drum, the user's emotional response should have an obvious link, via the system, with the user's actions. Measures of success might include the extent to which a user treats the system as if it is human. For example, does the system allow the user to "willingly suspend their disbelief" or not, and how often does a system get used voluntarily. In the context of HUMAINE's stated aim of trouble shooting in automated call handling systems, it is quite shocking just how much verbal abuse conversational agents receive (de Angeli 05) and the extent to which systems can "push back" as part of this affective loop is an open question.

Bibliography

Ask Jeeves, http://www.ask.com

Bates J (1994). The role of emotion in believable agents, Communications of the ACM, 37(7), 122-125.

Batliner, A., Hacker, C., Steidl, S. Nöth, E. Haas, J. (2004). From Emotion to Interaction: Lessons from Real Human-Machine-Dialogues. In E. Andre', L. Dybkiaer, W. Minker, P. Heisterkamp (eds.) Affective Dialogue Systems. Springer Verlag, Berlin, Germany, 2004

Beskow, J. Cerrato, L., Granstrom, B. House, D. Nordenberg, M. Nordstrand, M. Svanfeldt,G., (2004). Expressive Animated Agents for Affective Dialog Systems. In E. Andre', L. Dybkiaer, W. Minker, P. Heisterkamp (eds.) Affective Dialogue Systems. Springer Verlag, Berlin, Germany, 2004

Beskow J, Edlund JG and Nordstrand M (2005). A model for multi-modal dialogue system output applied to an animated talking head. In Minker, W, Bühler, D and Dybkjaer, L (eds) Spoken Multimodal Human-Computer Dialogue in Mobile Environments, Text, Speech and Language Technology, 29, Dordrecht, The Netherlands, Kluwer Academic Publishers.

Bosma, W. and Andre, E (2004). Exploiting Emotions to disambiguate Dialogue Acts, Proceedings of the 9th international conference on Intelligent User Interface, IUI 2004, Funchal, Madeira, ACM Press pp. 85-92.

Brahnam, Sheryl, Strategies for handling customer abuse of ECAs in:Abuse: the darker side of Human-Computer Interaction (INTERACT '05), edited by Antonella De Angeli, Sheryl Brahnam and Peter Wallis, Rome, September 2005.

Cavalluzzi, A., De Carolis, B., Carofiglio, V., Grassano, G (2003). Emotional Dialogs with an Embodied Agent. *User Modeling*: 86-95

Cavazza M., Martin O., Charles F., Mead S.J., Marichal X. and Nandi A., (2004). Multi-modal Acting in Mixed Reality Interactive Storytelling. IEEE Multimedia, July-September 2004, Vol. 11, Issue 3.

Cavazza, M. and Charles, F., (2005). Dialogue Generation in Character-based Interactive Storytelling. AAAI First Annual Artificial Intelligence and Interactive Digital Entertainment Conference, Marina del Rey, California, USA, AAAI Press.

Cahn JE (1990). The generation of affect in synthesised speech, journal of the American Voice Input/Output Society, 8, 1-19.

Cassell, J., Bickmore, T., Billinghurst, M., Campbell, L., Chang, K., Vilhjalmsson, H. and Yan, H. Embodiment in Conversational Interfaces: Rea in "Proceedings of the CHI'99 Conference", Pittsburgh, PA 520-527 1999

Cassell J, Sullivan J, Prevost S and Churchill E (2000). Embodied Conversational Agents, MIT Press Cambridge.

Colby (Kenneth Mark) Roger C. Parkinson and Bill Faught Pattern-Matching Rules for the Recognition of Natural Language Dialogue Expressions American Journal of Computational Linguistics, 1974

Cosi, P. Fusaro , A. Grigoletto, D. Tisato, G. (2004). Data-Driven Tools for Designing Talking Heads Exploiting Emotional Attitudes", In E. Andre', L. Dybkiaer, W. Minker, P. Heisterkamp (eds.) Affective Dialogue Systems. Springer Verlag, Berlin, Germany, pp. 101-112.

Cowie R, Douglas-Cowie E, Tsapatsoulis N, Votsis G, Kollias S, Fellenz W and Taylor J (2001). Emotion recognition in human-computer interaction. IEEE Signal Processing Magazine, 18, pp.32-80.

Cowie R. and Marc Schroder Piecing Together the Emotion Jigsaw in "MLMI 2004, LNCS 3361" edited by S. Bengio and H. Bourlard Springer-Verlag Berlin 305--317 http://emotion-research.net/aboutHUMAINE/ 2005

Damasio, A. R. Descartes' Error: Emotion, Reason and the Human Brain, Grosset/Putnam, New York. 1994

Darwin C (1872). The Expression of Emotions in Man and Animals, John Murray, London.

Davidson R, Scherer KR and Goldsmith H Eds. (2003). Handbook of Affective Sciences, Oxfrod University Press.

de Angeli (Antonella) Graham I. Johnson and Lynne Coventry The unfriendly user: exploring social reactions to

chatterbots in "Proceedings of The International Conference on Affective Human Factors Design" edited by Helander, Khalid and Tham, ASEAN Academic Press London, 2001

de Angeli (Antonella) Stupid Computer! Abuse and Social Identity in "Abuse: the darker side of Human-Computer Interaction (INTERACT '05)" edited by Antonella De Angeli and Sheryl Brahnam and Peter Wallis, Rome (http://www.agentabuse.org/) September, 2005

Douglas-Cowie E, Cowie R and Campbell N Eds. (2003). Speech and emotion, Speech Communication (special issue), 40(1-3).

Eisna, Roos Usable Technology for Older People: Inclusive and Appropriate (UTOPIA) http://www.computing.dundee.ac.uk/projects/UTOPIA/

Ekman P (1999). Basic emotions, Handbook of Cognition and Emotion, Dalgleish T and power M (Eds.), John Wiley, New York, 301-320.

Estival (Dominique) Spoken Dialogue for Virtual Advisers in a semi-immersive Command and Control Environment in "SIGdial Workshop on Discourse and Dialogue", Sapporo, Japan (citeseer.ifi.unizh.ch/646935.html) July 2003.

Gorin, A.L., Abella, A., Alonso, T. Riccardi, G. and Wright, J.H. (2002), Automated Natural Spoken Dialog, IEEE Computer Magazine, vol. 35 (4) pp. 51-56.

Gratch, J. (1999). Why you should buy an emotional planner. Proceedings of the Agents'99 Workshop on Emotion-based Agent Architectures (EBAA'99)

Gratch, J. and Marsella, S. Evaluating a Computational Model of Emotion Journal of Autonomous Agents and Multiagent Systems (Special Issue on the Best of AAMAS 2004) 2005

Haddad D et al (2002). Investigation and evaluation of voice stress analysis technology, US Dept. Justice, AFRL Report No.193832.

Henton C and Litwinowicz PC (1994). Saying and seeing it with feeling: techniques for synthesizing visible, emotional speech, Proc. 2nd ESCA/IEEE Workshop on Speech Synthesis, 73-76.

Iida, A., Campbell, N., Iga, S., Higuchi, F., and Yasumura, M. (2000) A Speech Synthesis System with Emotion for Assisting Communication, ISCA Workshop on Speech and Emotion, Belfast.

Johnson, W.L., Rizzo, P., Bosma, W., Kole, S., Ghijsen, M., van Welbergen, H. (2004) Generating Socially Appropriate Tutorial Dialog. In E. Andre', L. Dybkiaer, W. Minker, P. Heisterkamp (eds.) Affective Dialogue Systems. Springer Verlag, Berlin, Germany, 2004, pp. 254-264.

Johnson, W.L., Vilhjalmsson, H. and Samtani, P. (2005). The Tactical Language Training System (Technical Demonstration), , The First Conference on Artificial Intelligence and Interactive Digital Entertainment, Marina del Rey, CA, AAAI Press.

Kopp (Stefan), Lars Gesellensetter , Nicole Kramer and Ipke Wachsmuth A Conversational Agent as Museum Guide - Design and Evaluation of a Real-World Application in "5th International working conference on Intelligent Virtual Characters" (http://iva05.unipi.gr/index.html) 2005

Kreutel J. and Colin Matheson Modelling Dialogue using Multiple Inferences over Information States in "Proceedings of ICOS-2, 2nd Workshop on Inference in Computational Semantics", Dagstuhl, 2000

Mao, W. and Gratch, J. (2004). Social Judgment in Multiagent Interactions, in: Proceedings of the Third International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS 2004), New York, 2004.

Marsella, S. and Gratch, J. (2003). Modeling Coping Behavior in Virtual Humans: “Don't Worry, Be Happy," in: 2nd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2003), Melbourne, Australia, July 2003

Martinovski, B., Traum, D., and Marsella, S. (2005). Rejection of empathy and its linguistic manifestations in Procedings of Conference on Formal and Informal Negotiation, FINEXIN, Canada: Ottawa.

Mateas, M. and Stern, A. (2004). Natural Language Processing In Facade: Surface-text Processing. Technologies for Interactive Digital Storytelling and Entertainment (TIDSE 2004), Darmstadt, Germany.

Mei Si, Stacy C. Marsella and David V. Pynadath Thespian: Using Multi-Agent Fitting to Craft Interactive Drama in The Fourth International Joint Conference on Autonomous Agents and Multi Agent Systems (AAMAS 2005), Utrecht, The Netherlands.

Moore RK Ed. (1996). Speech under stress, Speech Communication (special issue), 20(1-2).

Muller, T.J., Hartholt, A., Marsella, S., Gratch, J., Traum, D.R. Do You Want To Talk About It? A First Step Towards Emotion Integrated Dialogue. In: E. Andre, L. Dybkjae r, W. Minker, P. Heisterkamp (Eds.), Affective Dialogue Systems: Tutorial and Research Workshop, ADS 2004, Kloster Irsee, Germany, June 14-16, 2004

Murray IR and Arnott JL (1993). Towards the simulation of emotion in synthetic speech: a review of the literature on human vocal emotion, J. Acoustical Society of America, 93, pp.1097-1108.

Nijholt, A. (2002). Embodied Agents: A New Impetus to Humor Research. In: The April Fools Day Workshop on

Computational Humour, O. Stock, C. Strapparava & A. Nijholt (eds.), Proceedings Twente Workshop on Language Technology 20 (TWLT 20), ISSN 0929-0672, ITC-IRST, Trento,Italy, pp. 101-111.

Oatley, K. and Johnson-Laird, P.N. (1987). Towards a Cognitive Theory of Emotions, Cognition and Emotion 1(1):29-50.

Ortony, A., Clore, G., Collins, A. (1988) The Cognitive Structure of the Emotions, Cambridge University Press, Cambridge.

Pelachaud C and Poggi I (2001). Towards believable interactive embodied agents, Fifth International Conference on Autonomous Agents workshop on Multimodal Communication and Context in Embodied Agents, Montreal.

Persson, P. EXMS: an animated and avatar-based messaging system for expressive peer communication, In Proceedings of the 2003 international ACM SIGGROUP conference on Supporting group work, pp. 31-39, Sanibel Island, Florida, USA. 2003.

Prendinger, H., Dohi, H., Wang, H., Mayer, S. and Ishizuka, M. (2004). Empathic Embodied Interfaces: Addressing Users' Affective State: Embodied Interfaces That Address Users' Physiological State. In: E. Andre, L. Dybkjae r, W. Minker, P. Heisterkamp (Eds.), Affective Dialogue Systems: Tutorial and Research Workshop, ADS 2004, Kloster Irsee, Germany, pp. 53-64

Picard, R. Affective Computing, MIT Press, Cambridge, MA. 1997

Rehm M and E. André. Informing the Design of Embodied Conversational Agents by Analyzing Multimodal Politeness Behaviors in Human-Human Communication. In: AISB Symposium for Conversational Informatics, 2005.

Riccardi, G. and Hakkani-Tür, D. (2005). Grounding Emotions in Human-Machine Conversational Systems. In: M. Maybury, O. Stock, W. Wahlster (Eds.) Intelligent Technologies for Interactive Entertainment: First International Conference, INTETAIN 2005, Madonna di Campiglio, Italy, Springer-Verlag, pp. 144-154.

Rigoll G, Müller R and Schuller B (2005). Speech emotion recognition exploiting acoustic and linguistic information sources, Proc. SPECOM, Patras, Greece, 61-67.

Scherer KR, Schorr A and Johnstone T Eds. (2001). Appraisal Processes in Emotion: Theory, Methods, Research, Oxford University Press, New York and Oxford.

Scherer KR (2003). Vocal communication of emotion: a review of research paradigms, Speech Communication 40, 227-256.

Schröder M (2001). Emotional speech synthesis: a review, Proc. Eurospeech, Allborg, Denmark, 561-564.

Slater, M., Steed, H. A. and Gaurau, P. M. (2000) Acting in Virtual Reality, In Proceedings of the third international conference on Collaborative virtual environments, pp. 103-110, San Francisco, California, United States.

Streit, M., Batliner, A. and Portele, T (2004). Cognitive-Model-Based Interpretation of Emotions in a Multi-modal Dialog System. In: E. Andre, L. Dybkjae r, W. Minker, P. Heisterkamp (Eds.), Affective Dialogue Systems: Tutorial and Research Workshop, ADS 2004, Kloster Irsee, Germany, June 14-16, 2004, pp. 65-76.

Sundström (Petra) Exploring the Affective Loop, PhLic Thesis, Stockholm University, Sweden, 2005.

Turunen (Markku) Esa-Pekka Salonen, Mikko Hartikainen, Jaakko Hakulinen, Kristiina Jokinen, Jyrki Rissanene, Kari Kanto, Antti Kerminen, William Blacm, Allan Ramsay, Adam Funk, Andrew Conroy, Paul Thompson, Mark Stairmand, Björn Gambäck, Magnus Sahlgren, Fredrik Olssom, Maria Cheadle, Preben Hansen, and Stina Nylander AthosMail: a Multilingual Adaptive Spoken Dialogue System for the E-mail Domain in "Robust and Adaptive Information Processing for Mobile Speech Interfaces (DUMAS Final Workshop)" edited by Björn Gambäck and Kristiina Jokinen (http://citeseer.ist.psu.edu/729314.html) August, 2004

Walker, M., Cahn, J., and Whittaker, S. (1997). Improvising Linguistic Style: Social and Affective Bases for Agent Personality . In Proceedings of the Conference on Autonomous Agents, AGENTS97.

Wallis, P., Helen Mitchard, Damian O'Dea and Jyotsna Das, Dialog Modelling for a Conversational Agent in "AI2001: Advances in Artificial Intelligence, 14th Australian Joint Conference on Artificial Intelligence" edited by Markus Stumptner, Dan Corbett and Mike Brooks, Adelaide Australia, Springer (LNAI 2256) 2001

Weizenbaum (Joseph) ELIZA - A Computer Program For the Study of Natural Language Communication Between Man and Machine Communications of the ACM vol 9(1) January 1966

Wilks, Y. (2004) Companions : a new paradigm for agents. Proc. International AMI Workshop, IDIAP, Martigny,CH.

Wundt W (1874). Grundzüge der Physiologischen Psychologie, Engelmann, Leipzig.

Yildirim S et al (2004). An acoustic study of emotions expressed in speech', Proc. InterSpeech, Jeju, Korea.

Zukerman (Ingrid) and Richard McConachy WISHFUL: A Discourse Planning System that Considers a User's Inferences Computational Intelligence vol 17(1) Blackwell Publishers Boston MA & Oxford UK, Februrary 2001