

1 Introduction and Overview

1.1 Introduction

There is plenty of evidence that social behaviour plays a very significant role when people interact with their computers. There are studies showing that people impose their interpersonal behaviour patterns on their computers [Reeves and Nass, 1996], and that people *expect* their computers to interact in a human-like way [Nass et al., 1994, Nass and Sundar, 1994]. Humans even perceive computers to have personalities [Moon and Nass, 1996, Dryer, 1999]. And that's *before* any attempt has been made to make computers behave more like humans.

It is taken as read in the field of human-computer interaction that there are interactions between humans and their computers which are subjectively and objectively improved when computers are equipped with programs that make them interact in a more human-like fashion, and that the ability to interact with '*emotional intelligence*' is particularly important in this context. After all, emotions affect how people remember, how they learn, and how they make decisions [Picard, 1997]. Many examples of specific improvements have been given in the literature: companies will lose fewer customers if the computer program they interact with can tell when the customer becomes upset or frustrated and take appropriate action [Batliner et al., 2000, Ang et al., 2002, Devillers et al., 2002]; in a situation where users receive bad news through a computer program they will be a lot happier if the program can do this with apparent empathy and sympathy [de Rosis and Grasso, 2000]; internet chatting becomes more personal and enjoyable if appropriate emotion tags are displayed along with the text [Holzman and Pottenger, 2003]; troubleshooters are more efficient if they can adapt to users' emotional state and personality [Breese and Ball, 1998]; the effectiveness of an English tutor in a simulated conversation set-up is improved if the tutor is socially and emotionally 'aware' [Prendinger and Ishizuka, 2001]; and if that's not enough, [Picard, 1995] has another 50+ examples. Inevitably, there is a potential intelligence and/or military use, and research in this area has e.g. been carried out as part of a "chat-mining research project" for the U.S. government/military [Holzman and Pottenger, 2003], and military training simulators [Gratch, 2000, Fleischman and Hovy, 2002].

The most obvious way in which computers can be made to appear emotional is by communicating in emotional ways. There are two sides to emotional communication — the ability to express emotions and the ability to recognize emotions in others. Emotion can be expressed in ways that can be *seen* (facial expression, gestures, moving about etc.), and in ways that can be *heard* (prosody and other vocal characteristics). However, the research surveyed in this paper is concerned with how emotion is expressed *verbally*, that is to say in the word string, or, more precisely, the part of language that can be orthographically transcribed.

The main focus of this survey is emotion-dependent variation of automatically generated language, or **Emotional NLG** for short. Although no longer young, this is still an emerging field of research that concerns itself with varying language output (in NLG systems) to reflect different emotions. However, the survey also includes a literature review of the closely related (and similarly small) field of **Verbal Emotion Identification** which develops methods for spotting emotional signifiers in language. The reason for the inclusion is that there is substantial overlap between the two fields: (i) both seek to identify lexical, syntactic and semantic properties that are correlated with specific emotions; (ii) both have employed psychologically inspired models of emotion. Furthermore, as will be argued, emotion identification methodology can provide quantitative, empirical grounding for emotional NLG.

1.2 Concepts and Terminology

Realism and believability: When researchers on computer-based characters talk about what it is they're trying to do, they often use words like *realistic*, *human-like*, *life-like* and *natural*. E.g. the ITRI NECA homepage says the aim of the project is to create "on-line beings which are able to speak and act like humans." But that's not really the point at all, not least because it implies that artificial characters are better the closer they are to real humans. Artificial characters have their own terms of plausibility. People have been creating characters for as long as there has been any form of fiction, and audiences don't have difficulty deciding whether a character works for them, i.e. whether they find it plausible. So in this respect the task of the computer-based character designer is just the same as that of any creator of written or visual fiction. This view is shared by many working on computer-based characters, and is reflected in the frequent use of the term *believable* to describe what computer-based characters should be like.

Computer-based characters: The many terms that have been used for entities of this kind include *artificial agents*, *avatars*, and *talking heads*, but I think the most sober, clearest and most generic is *computer-based characters*.

Emotional entity: Where the reference is to human emotion only I have tried to make that clear, but there are many cases where I say something about emotion that isn't specific to computer-based characters or humans, in which case I've opted for the non-committal term *emotional entity*.

Signifiers of emotion: This term is intended to cover what has been called *signals, signs, cues* and *correlates of emotion* (doubtless among many other terms), that is to say, small semantic units with emotional content (e.g. frown, head-shake, swear word, an interjection, a particular syntactic disfluency, etc.).

Verbal expression of emotion: The non-auditory, non-visual part of emotion expression through language, that is, emotion expressed in the word string.

Emotion-dependent variation in NLG: Emotion-dependent variation in language means variation between utterances that can be accounted for exclusively by a specific emotional state. To use the linguistic concept of minimal pairs: An NLG system implements *emotion-dependent language variation* if it is capable of generating minimal emotional pairs ‘at the flick of a switch’ (where the switch is not necessarily controllable from the outside). A minimal emotional pair is a pair of utterances such that the difference between them can be explained entirely in terms of emotion.

Emotion: In terms of developing a framework for computationally modelling and expressing emotion, I don’t think it matters what exactly different people would include under the heading of *emotion*, as long as the framework can accommodate a broad range of definitions and implementations.

1.3 Overview

The report starts with a survey of literature on emotion in NLG (Section 2) and on what can be seen as the inverse — automatic emotion identification (Section 3). In these surveys, the focus is on generation and recognition of emotional variation in the word string, as opposed to speech, gestures or physiological signs. The same focus underlies the annotated bibliographies that follow (Section 4). The decision to include only those publications that report some work involving properties of word strings has meant the the number of papers included is rather small. The report continues with annotated lists of research groups and research projects with a focus on emotional NLG, and of data collections that are in some way annotated for emotion (Section 5). It closes with a small set of conclusions in Section 6.

2 Emotional Variation in NLG

There are very few reported pieces of research that have developed a way of varying language depending on emotions in the sense defined above: an early example is Klein et al.’s work on the Automatic Novel Writer (ANW) [Klein et al., 1973], then there is Hovy’s thesis research [Hovy, 1986, Hovy, 1988, Hovy, 1990], and Fleischman and Hovy’s (2002) work for the MRE system, where ANW and MRE are on a much smaller scale.

In the Automatic Novel Writer, some relations could be modified by events in the simulation, e.g. the relation “AFFECTION” had a numeric variable parameter (5=love, 4=like, 3=know, 2=dislike, 1=hate) that determined lexical expression. These values were used to describe relations between characters, and they could be incremented or decremented by various simulated events.

PAULINE’s fine-grained and sophisticated approach to generating text that varies in accordance with the interlocutors’ goals and attitudes covers content selection and realisation, and models both the hearer’s and the speaker’s goals and attitudes. However, it is in large parts based on a huge number of very complex rules that decide when to activate goals and strategies. All of this is implemented as an interleaved planning and realisation regime. This would make it very difficult to maintain and extend the system, and it’s difficult to see how interaction between rules and therefore undesirable side effects can be predicted.

The MRE system is by comparison exceedingly simple. Only the speaker’s attitudes are taken into account and only realisation is covered (while the system can choose to leave out information about an event’s agent and/or patient present in the semantic representation, it can’t add/leave out larger chunks of information, and it doesn’t have the ability to add any information at all on the basis of affective considerations). This limits the degree to which utterances vary. A large number of realisations are considered, but it’s still only a small proportion of all realisations consistent with the semantic input, as variation is based exclusively around verb frames.

There are a few other, closely related pieces of research. Walker, Cahn & Whittaker (1996, 1997) created a framework for social interaction where speech act execution strategies are selected on the basis of social factors, but emotions only figure directly in that an agent’s emotional disposition is set and all its utterances are then synthesized with the acoustic signifiers for that emotion.

Loyall & Bates (1997) while describing a system which is in principle suitable for making language generation dependent on emotion, did not actually implement such a thing. Nothing was done in this direction in the rest of the Oz Project either.

In the MagiCster Project, emotion was also not directly coupled with variation in the verbal output. From De Carolis et al. (2002) it seems that the emotive wording in doctor-patient dialogues results from choice of dialogue (sub)plan rather than emotion as determined by the Affective Agent Modelling component.

A lot of project descriptions and abstracts of papers make it sound as if a particular system/piece of research bases choice in NLG on emotion but more often than not it turns out to be a theoretical possibility rather than an implemented reality. E.g. despite publications apparently claiming the opposite, selection of content, syntactic structure and lexical items is not connected *at all* to an agent's emotional state anywhere in the long-running Oz Project¹.

A lot of research gets close to connecting emotion and NLG but boils down to preparing the ground for the connection. Many look at directly related issues: mood, personality, other variables such as the social distances between speakers, the power of the hearer over the speaker, and the 'basic desires' of speaker and hearer [Walker et al., 1997]. Other indirectly related work includes:

- modelling social structure and how it influences selection of basic speech act execution strategies (but syntactic and lexical choice is not covered) [Walker et al., 1997]
- emotional speech synthesis [Cahn, 1990]; see also the overview in [Schröder, 2001]
- reconciling emotions with other aspects of human behaviour: integrating emotion with actions or behaviours [Loyall and Bates, 1997]; integrating reasoning about emotions and plans [Gratch, 2000]
- a long list of publications about believable agents, mostly agreeing about why they're a good idea, and that they need to have personality and emotion e.g. [André et al., 2000]
- affective reasoning [Ortony et al., 1988, Elliot, 1992, Prendinger and Ishizuka, 2001];

Among the papers that directly discuss emotion-dependent variation in NLP there is a focus on realisation, but very little on content planning. Exceptions are Hovy's thesis, and [Kölln, 1995] who incorporate subjective preferences towards domain concepts during content selection.

In some cases, approaches to emotion modelling and emotion generation are based on researchers' intuitions, in others on existing psychological and cognitive theories of (human) emotion. There are no reported approaches based on more quantitative, empirical analyses (even given a much more loosely defined notion of emotional NLG), presumably because corpora on emotion are hard to come by (and hard to develop in the first place).

A number of specific (and unconnected) generation strategies have become popular, e.g. concerning sentence structure: don't aggregate under certain circumstances, e.g. when user is emotional [Walker, 1992, Walker, 1996, de Rosis and Grasso, 2000]. A related intuition concerns content selection: e.g. [Walker, 1996] mention that there are many reasons why sometimes the normally useful rule "don't tell hearer what he/she already knows, tell them what they don't know" shouldn't be applied. Similarly, [de Rosis and Grasso, 2000] point out that if the hearer is likely to be upset (their example is doctor giving patient diagnosis) then there is a need to over-explain because hearer isn't likely to be listening too carefully.

Emotion modelling for NLG has been kept very simple: in the Automatic Novel Writer, the affection relation had values ranging from 1 to 5 that could be incremented or decremented by the occurrence of certain events. In Hovy's PAULINE emotions are basically modelled as affect values (one of good, bad or neutral) that influence goal selection strategies, content selection and word choice. Fleischman & Hovy have values from -5 to 5 attached to each event, agent, patient, etc. "representing how positively or negatively the speaker feels toward the element" (p. 59).

In contrast to emotion identification (see following section) emotional NLG is extremely hard to evaluate, and there is nothing but the most informal, small-scale tests.

3 Emotion Identification

This research area is also known as *affect sensing*, *sentiment classification* and *emotion recognition*. It is concerned with the task of deciding whether emotion is being expressed in a given piece of language, and usually also what kind of emotion it is. This is invariably construed as a classification task, where utterances are mapped to emotion categories.

As with Emotion Generation, much of the research on emotion identification has focussed on auditory (prosodic) cues, but linguistic features (particularly lexical ones) have also played a role.

The overall trend has been to keep the task definition simple, and the number of classes small (often only two emotions, or an emotional category vs. a neutral category). Representative numbers are e.g. 8 emotions (Vyzas and Picard, 1999) when using physiological cues, 3 emotions [Whiteside, 1998, Li and Zhao, 1998] for speech cues, and 2

¹<http://www-2.cs.cmu.edu/afs/cs.cmu.edu/project/oz/web/oz.html>

[Pang et al., 2002, Ang et al., 2002] when using only text. Moreover, results tend to be reported for artificial scenarios, experienced speakers and elicited controlled speech [Batliner et al., 2000].

Of course, automatic identification rates depend on how complex the task definition is, and what kind of cues are used. The most reliable are *physiological cues*, where automatic identification rates above 80% and even above 90% can be achieved [Vyzas and Picard, 1999, Healey, 2000], respectively.

Humans do well identifying emotion from *facial expression* achieving around 70% (University of Pennsylvania face rating task), compared to accuracies as high as 90% for automatic methods [Bartlett et al., 1999].

In emotion identification from *speech*, humans perform similarly, e.g. [Polzin and Waibel, 2000] report 70% correct recognition for three classes. Automatic methods achieve results as high as 90% for binary distinction tasks [Zhou et al., 1998], decreasing to 60–80% for slightly less simple task definitions (e.g. happiness, sadness/neutral and anger [Whiteside, 1998, Li and Zhao, 1998]), and further decreasing for larger sets of emotion categories. Among the possible cues from speech, spectral information (pitch, intensity, voice quality) is a more reliable basis for emotion identification than prosody (fundamental frequency, etc.) [Polzin and Waibel, 2000].

Information extractable from the word string generally has been found to be a less successful basis for recognising emotions. Humans do consistently worse identifying emotions from text only as compared to speech, e.g. [Polzin and Waibel, 2000] report 70% correct from speech, and 55% from text for sad/angry/neutral. Automatic methods typically achieve results in the mid-60s for text.

[Ang et al., 2002] create two language models (class-based trigram models), one each for their two classes (frustration/annoyance and everything else), then compute the sign of the difference between the log likelihoods of a given utterance according to the models (intuitively, one sign can be interpreted as a decision in favour of one class, the other as a decision for the other class). This decision mechanism on its own achieved an accuracy in the mid-60s.

[Devillers et al., 2002] build five emotion models (unigram models), one for each of their emotion classes. The model for ‘neutral’ is trained on the entire training data (which includes emotional utterances) and is interpreted as the ‘general task-specific model’. The approach selects the emotion category for which $\log P(\text{utterance}|\text{Emotion})$ is highest. The basic approach achieves an overall precision of 61.6%. Stemming improved this to 64.8%, additional compounding (26 expressions) to 67.2%.

Among the features that have been used in emotion identification from the word string are the following:

- position of utterance in dialogue [Ang et al., 2002]
- whether utterance is a repeat/rephrase/correction [Batliner et al., 2000, Ang et al., 2002]
- likelihood of utterance according to different language models (modelling the likelihoods of sequences of words): e.g. class-based trigram model [Ang et al., 2002], word-based unigram model (with stemming and compounding) [Devillers et al., 2002]
- keywords: (i) straightforward keyword spotting — e.g. [Elliot, 1992] based on OCC model and Affective Lexicon [Ortony et al., 1988]; (ii) spotting emotionally salient words [Lee et al., 2002]
- type of dialogue act / sequences of dialogue acts [Batliner et al., 2000] (intersects with repeat/rephrase/corrections)
- also: Latent Semantic Analysis [Landauer and Dumais, 1997]: popular for affect classification of text e.g. Baby Webmind [Goertzel et al., 2000]; drawback: works better on larger chunks of text

If used without other types of information (such as prosodic), verbal features (which usually means lexical ones) often achieve poor results for emotion identification. Straightforward keyword spotting is often criticised for its apparently obvious flaws such as the fact that it can’t account for negation, and that it doesn’t have access to the true meaning of utterances [Liu et al., 2003, p. 127]. However, it turns out that lexical items, if used in the right way, are a powerful predictor of emotionality. [Pang et al., 2002] achieve a best result of 82.9% accuracy for the task of classifying film reviews into positive and negative, where each review is represented as a binary vector with 16,165 bits where each bit encodes presence or absence of a word (without stemming or stoplists!). Classification was done with support vector machines. Because they don’t do any *a priori* selection of keywords, their result implies that poorer results may be due to bad selection methods.

Some of the publications mentioned in this section are described in more detail in the following section. The following table summarises results for verbal emotion identification:

publication	best result	emotion categories	method	data
[Ang et al., 2002]	65.6% 69.8% 80.2%	2: annoyed/frustrated vs neutral/tired/amused/other = =	language model repeat/correction language model, repeat & speech features	natural human/machine = =
[Lee et al., 2002]	70% 75.65% 80.75%	2: neg/non-neg = =	acoustic features & linear discr analysis linguistic features combination of above	real human/machine dialog (female only) = =
[Pang et al., 2002]	82.9%	2: pos/neg	word presence	film reviews
[Polzin & Waibel, 2000]	46.7% 60.4% 63.9%	3: sad/angry/neutral = =	prosodic model spectral model bigram language model	movie speech = =
[Devillers et al., 2002]	67.2%	5: anger, fear, satisf., excuse, neutral	unigram models, stemming & compounding	real dialogues

4 Annotated Bibliographies

4.1 NLG with emotions

[Carolis et al., 2002] *From Discourse Plans to Believable Behavior Generation:*

This is a report on how (in MagiCster and using the Festival speech synthesizer) “a typical NLG architecture has been changed to generate context-adapted behaviour in a conversational embodied agent” (p. 65).

The strategy for planning the agent’s behaviour is for the planner to decide on the discourse steps required to achieve a given communicative goal. Two possibilities are considered in the discussion: (i) the planner decides what verbal/nonverbal ‘signals’ to employ; (ii) the planner decides communicative function(s) only, and the surface realizer decides the signals. Arguing in favour of strict mind/body separation, the authors opt for (ii): the agent’s mind decides what to say, the body interprets and renders this at the surface level along the different available channels (different agents have different body capabilities). In the first phase, a “plan library” is used — a collection of recipes for (i) plans that represent the steps required to achieve a given communicative goal, and (ii) the discourse plans that correspond to each of the steps in (i). A ‘Plan Enricher’ is used as the interface between the two phases. Within Plan Enricher, a transformation algorithm called Midas transforms input DPML trees into output APML (markup language for behaviour specification) trees. Midas recurses through the tree top-down until it reaches a leaf-node at which point the ‘generate_performative’ function is called: “This function is responsible for the surface realisation, in which the <performative> element is generated. If the Affective Agent Modelling component establishes that an emotion is felt by the Agent [...] and that this emotion has to be displayed, the affect attribute of the performative tag is set to that emotion’s name. [paragraph] The generate_performative function [also] produces the verbal part of the speech act [...]” (p. 71) (which implies that there can only be one emotion, and that all the text generation is done within the plan enricher).

What is not explained is how the ‘verbal part of the speech act’ is selected and whether the selection process is conditioned on the emotion tag. This is not very likely since the same function attaches the tag and selects the realisation. It appears that the apparently emotive phrasing comes directly from the earlier selection of a discourse plan, e.g. the plan for ‘doctor talking to patient’.

Additional information from the MagiCster Deliverable 2.3 (31/10/1002): “The generation component in Prototype 1 uses a relatively straightforward template-based system developed by DFKI. [...] The disadvantage [of such approaches] is that it is difficult to introduce flexibility into the actual language which is generated.” (p. 14)

[de Rosi and Grasso, 2000] *Affective Natural Language Generation:*

This is a first-principles piece about where and how to integrate affect in NLG. The paper looks at how NLG should vary according to the hearer’s personality, current emotional state etc. (e.g. when doctor has to deliver bad news to patient). The paper is not about modelling how an artificial communicating entity expresses its own emotions. There isn’t much of an empirical basis, decisions on what to do in discourse planning, sentence planning, and especially realisation have a somewhat ad hoc feel to them: “Affective text may be obtained by employing rule-based heuristics that define when and how empathy elements have to be introduced in the text.” (p. 213) The paper doesn’t describe any implementation.

[Fleischman and Hovy, 2002] *Towards Emotional Variation in Speech-based Natural Language Generation:*

This paper describes the MRE generator which is implemented in the SOAR programming language (Newell, 1990), and has three modules: sentence planning, realisation and ranking. This paper deals with the conversation mode of the system only.

The sentence planner gets as input minimal information about a state/event to be described with references to the actors and objects involved in it. A set of SOAR production rules converts this into an enriched case frame structure. This conversion relies heavily on the emotional decision engine.

The realizer is highly lexicalised. Tree construction begins with verb selection. Each verb has slots for its constituents (e.g. agent/patient). All possible realisations are produced.

The ranker determines each tree's score based on the tree's information content and emotional quality, where the score for the tree is derived from the sum of the scores for the nodes.

Emotional language generation is viewed as an optimization problem, minimizing the distance between the emotion of the speaker and the emotion of the sentence.

The emotion model is based on appraisal theories of emotion (OCC, Lazarus) and construal frames from Clark Elliot which represent relations between events and agents' dispositions (plans and goals in MRE). "[...] an agent's emotional state is predicated entirely on that agent's appraisal of an event in terms of how that event relates to its own set of goals, and plans toward those goals." (p. 59)

The emotional state of the speaker towards each element of its world model is an integer value v , $-5 \leq v \leq 5$ calculated by emotion model. During planning, all objects are assigned a frame by selecting the semantic frame with the minimal distance between the frame's default emotional value and the emotional attitude of the speaker towards the hearer.

In realization, all verbs in the lexicon that are valid representations of the input frame are used to create distinct trees. Lexicon entries have three emotion shades for each verb: event, agent, patient. The tree is then selected "in which the *total* emotional distance from the speaker's attitude is minimized across the event itself, as well as across all the constituents of that event" (because these can differ).

This approach is strictly verb based, permitted realisations include those that elide agent or patient. The total score takes into account both the degree to which the semantic content is expressed *and* the emotional distance between realisation and speaker.

[Haimowitz, 1991] *Modeling all Dialogue System Participants to generate Empathetic Responses:*

According to De Rosis and Grasso (2000), this paper "focuses on the idea of "producing utterances empathetic to both the Speaker and the Hearer", by offsetting "unpleasant" information and stressing "favourable" info, through densifier and intensifier adverbs: the Hearer's mental model is enriched, to this purpose, with domain-related personal preferences, concerns, worries and related features."

[Hovy, 1990] *Pragmatics and Natural Language Generation, Section 6 "Partiality":*

This section provides a summary of the approach to affect in PAULINE. It distinguishes biasing content and form — there are rules for biased topic selection, and there are rules for word choice. The paper looks generally at "why and how [it is] that we say the same thing differently to different people, or even to the same person in different circumstances".

Two main findings: "any generation system sophisticated enough to operate in service of communicative goals will have to maintain an intermediate level of goals and strategies, called here rhetorical goals." and: "any generation program flexible enough to operate under a number of communicative goals [...] will have to monitor the effects of its individual utterance components under an interleaved planning-realization regime."

Hovy admits that PAULINE solves each of the questions it addresses "by a set of simplified, somewhat ad hoc methods".

Models are based on available time, tone of interaction, speaker's opinion about the subject, depth of acquaintance between interlocutors, whether there is a goal to influence hearer's opinion, affect for the topic, desired effect on hearer's emotion toward speaker, desired effect on interpersonal distance, speaker-hearer relative social status, desire to involve hearer, knowledge level, desire to involve speaker.

Affect in PAULINE is modelled as $affect = v, v \in good, bad, neutral$. Values come from two sources: (i) provided by user (specifies one or more representation elements as sympathies or antipathies), (ii) defined as intrinsic to certain representation elements (e.g. in neutral context, concept 'arrest' is bad). PAULINE can combine and propagate different affects.

Inputs are one or more initial sentence topics. PAULINE then uses one of three topic collection plans (convince, describe, relate) to collect additional topics from the concept representation network.

The rhetorical goal is defined as $partiality = v, v \in partial_explicit, partial_implicit, impartial$. There is a set of rules that decide between partial/impartial and explicit/implicit, conditional on such things as affects towards the topic and towards the other speaker.

PAULINE uses style strategies to decide at choice points. E.g. in the case of explicit partiality, the strategy is to include explicit expressions of opinion. In the case of both explicit and implicit partiality, affective adjectives and

adverbs as well as stress words are included (in the sentence constituent inclusion component), and nouns and verbs that carry affect are selected in word choice component.

[Loyall and Bates, 1997] *Personality-Rich Believable Agents That Use Language:*

Loyall and Bates describe the details of an NLG extension to Hap, “the behavior-based architecture used by the Oz group to construct non-linguistic believable agents”. This extension is based on Kantrowitz’s GLINDA.

The paper outlines some of the things that become possible once language generation and behaviour generation are integrated in Hap, but none of this appears to have been implemented.

There’s a description of how language generation can be integrated with emotion in principle, but the paper doesn’t reveal what has been implemented, and there’s nothing that amounts to a theoretical approach to the integration (just an outline of how Hap permits *some* approach to be implemented). The paper refers to Kantrowitz’s research being about exploring pragmatic variation for natural referring expressions, but there don’t seem to be any publications on this.

See also entry for Oz Project in Section 5.

[Walker et al., 1997] *Improvising Linguistic Style: Social and Affective Bases of Agent Personality:*

This paper is about LSI (linguistic style improvisation) which is based on Speech Acts theory (James Allen, Philip Cohen, 1970s), and social anthropology and linguistics research on social interaction, in particular Brown and Levinson’s theory of linguistic social interaction.

LSI is a (partial) implementation of Brown and Levinson’s (1987) theory of social interaction (L&B), which comprises concepts (agents have ‘face’ consisting of desire for autonomy and desire for approval), capabilities for rational reasoning, and social variables (social distance between speaker and hearer, power of hearer over speaker, imposition ranking).

The approach uses a set of initiating speech acts: inform, offer, request-info, request-act; and a set of response speech acts: accept-inform, accept-offer, accept-request, reject-inform, reject-offer, reject-request. Speech act definitions include (a) the conditions under which speaker achieves his communicative intention, and (b) the effects on the hearer if the speaker succeeds.

The approach follows earlier work in that speech acts are implemented in a standard AI planning system (e.g. Litman and Allen, 1990); each speech act definition has preconditions and effects, as well as several ‘decompositions’ which specify the different ways in which the speech act can be realised.

The speaker calculates threat theta to the hearer resulting from a speech act as the sum of the social distance, power and imposition ranking. Theta is then the basis for choosing a speech act execution strategy, where choices range from direct execution to hinting and/or making utterance ambiguous (interestingly, they leave out B&L’s 5th possibility of not executing a speech act at all). Each strategy is then executed by a range of substrategies whose semantic content is selected from the plan-based representation for the speech act, and whose syntactic form is selected from a library (but the selection mechanism not clear). There is no grammar as such, just a “library of syntactic forms” (p. 99).

The emotional disposition of each agent is set (one of angry, annoyed, disgusted, distraught, gruff, pleasant, sad) and all of the agent’s utterances are synthesized with the acoustic correlates of that emotion (based on Cahn’s (1990) theory of expressing affect in synthesized speech and using her Affect Editor program).

Social structure, and (apparently) emotion remain fixed during dialogue.

As far as I can see *only* acoustic variation depends on emotions, not word strings.

This work seems to have petered out - there seem to be no further publications, and the 96 paper isn’t mentioned or cited anywhere.

4.2 Emotion identification

[Ang et al., 2002] *Prosody-based Automatic Detection of Annoyance and Frustration in Human-Computer Dialog:*

Ang et al. report research carried out as part of the DARPA Communicator Project. Detection of annoyance and frustration is construed as a classification task on utterances. The classifier used is a decision tree, with features reduced to a minimum by a brute force selection algorithm. Apart from numerous fine-grained prosodic features, three non-prosodic ones were used:

- a feature representing the position of the utterance in the dialogue
- a feature encoding whether the utterance is a repeated request or correction, with the following values: ‘not a repeat/correction’, ‘repeat or rephrase only’, ‘repeat or rephrase with explicit correction’, ‘explicit correction only’
- a language model feature: the value is obtained by computing two log likelihoods of the utterance with two different language models (class-based trigram models), then taking the sign of the difference.

There were two binary classification tasks. In the first task, one class was ‘annoyed or frustrated’, while the other class was ‘neutral, tired, amused and any other’. The second task was the same as the first except that annoyance was removed from the first class. For task 1, human accuracy (measured as agreement with the consensus) was 83.9%, compared to 80.2% achieved by the automatic classifier using all features. For task 2, human accuracy was 77.3% compared with 93.2% classifier accuracy (all features). For task 1, The language-model feature on its own achieved an accuracy of 65.6% on the complete data, compared to 69.8% achieved by the repeat/correction feature on its own. No results are presented for position only.

[Batliner et al., 2000] *Desperately Seeking Emotions: Actors, Wizards, and Human Beings:*

This paper reports work from the SmartKom Project. The problem considered is a two-class classification problem with the emotion classes neutral and anger. Three different experimental settings are investigated: (1) speaker-specific with data from one experienced speaker acting; (2) speaker-independent with data from several inexperienced speakers reading; and (3) speaker-independent with data from several naive speakers obtained in a Wizard-of-Oz scenario. All data was annotated emotional or nonemotional.

In setting (3), 20 dialogues (2,395 turns) were recorded of naive users interacting with a ‘malfunctioning’ Verbmobil system. The idea was that exactly the same situation is repeated many times with the user becoming increasingly angry, so that the change in linguistic properties within an unchanging setting can then be observed over time. Annotation included three types of ‘peculiarities’: conversational (cooperative, insulting etc.), prosodic (very clear articulation, pauses between words), and lexical (swear words etc.). All turns annotated with more than one peculiarity were annotated emotional (otherwise nonemotional).

Classification was done with three different classifiers. Results for multi-layered neural networks using r-prop as a training algorithm are presented but not discussed. The two remaining classifiers are cart and regression tree classifiers and linear discriminant analysis, for both of which leave-1-out cross-validation was carried out. Prosodic feature vectors formed the inputs, with additional segmental information in some tests. The authors report results that decrease from setting 1 to 3, as expected because elicited emotion is more clearly expressed than naturally observed emotion. Best results were achieved for linear discriminant analysis using both acoustic and segmental information: 97%, 82% and 71% correctly classified turns for settings 1, 2, and 3 respectively.

The main result appears to be that prosody isn’t good enough to detect emotion in real-life communication. The conclusion is that other information needs to be added, and the paper finishes with a preliminary report on work in progress. The system MoUSE has 6 classifiers to spot “marked behaviour”, of which three have been implemented: prosody, repetitions, reformulations). The preliminary (and obvious) result is that when combining information about repetition, language peculiarities (as above) *and* prosody into account, then spotting “trouble in communication” becomes easier.

[Devillers et al., 2002] *Annotation and Detection of Emotion in a Task-oriented Human-Human Dialog Corpus:*

This paper reports work carried out within the IST Amities Project. The data that was used is a corpus of 100 human (agent-client) dialogues recorded in a stock exchange service center which are annotated with labels of three types: dialogue, dialogue progression and emotion. The authors make much of using naturally occurring speech, but the dialogues aren’t ‘real’ because clients aren’t real clients (emotions would probably run higher if people were talking about their own investments).

The first study looks at correlations between emotions and labels marking dialogue quality, progression and success. Correlation tables show mostly weak correlations between dialogue and emotion features. Linear regression was used to find combinations of features with high predictive accuracy for some individual emotions. For the two emotions for which results are reported (anger and fear), the predictive accuracy was low (48.6% for anger, 30.6% for fear).

In a second study, a model for emotion detection was built based on the hypothesis that there are types of lexical information particularly salient for different emotions. Corpus sentences were reannotated for emotions by new annotators who looked only at the randomly ordered orthographic transcriptions. In the text-based annotation, anger and fear labels decreased a lot, sadness, excuse and neutral increased (as compared to the annotation from recordings of speech).

Five emotion models (unigram models) were built. The model for ‘neutral’ was trained on entire training data (including utterances tagged for emotion) and was interpreted as the “general task-specific model”. The classification approach selects the emotion for which $\log P(\text{utterance} | \text{Emotion})$ is highest.

This basic approach achieved an overall precision of 61.6%. Stemming improved this to 64.8%, additional compounding (26 expressions) to 67.2%. Using lists of words to exclude from the models didn’t improve results. Normally the n most frequent are used, but in this work it’s an ad hoc selection of words likely not to matter. There are significant differences between detection rates for different emotion categories: from 38% for fear to 88% for neutral and satisfied.

Individual words that were found to be particularly reliable identifiers for categories:

anger:	swear words, ‘abnormal’, ‘irritating’, ‘embarrassing’, ‘bothering’
fear:	‘worry’, ‘fear’, ‘panic’, ‘afraid’, ‘disastrous’
satisfaction:	‘agree’, ‘thanks’, ‘perfect’, ‘excellent’
excuse:	‘mistake’, ‘error’, ‘sorry’, ‘excuse’, ‘pardon’

A third study was carried out to look at the difference in human judgments when listening to audio as opposed to reading the transcriptions. Only 55% of the utterances were labelled the same under both conditions. Surprisingly, in the labelling from transcriptions, the proportion of utterances labelled emotional is much *higher* than in the labelling from audio.

Subjects were asked to identify the prosodic (forced choice) and lexical cues that helped them make their judgments. Results indicate that subjects used fast speech for irritation and satisfaction; a flat F0 for neutral and excuse; a variable F0 for emotional; and gained no help from intensity. Keywords subjects tended to use were words that give away the emotion (nervous, i’m afraid...), swear words, exclamations, and negations. There were no definite results on whether subjects used any syntactic characteristics.

Interestingly, although this information can only be gleaned from Figure 1, anger was not identified at all from the audio version, but relatively well from the text-only version.

Another interesting observation is that there is evidence that there can be a contradiction between prosodic information about emotion and the words of the utterance (Section 5.2).

A peculiarity of the paper is not presenting separate results for agents (who are much more likely to control/hide emotions) and clients.

[Lee et al., 2002] *Combining Acoustic and Language Information for Emotion Recognition:*

This paper reports research incremental on the authors’ 2001 paper [Lee et al., 2001], the main difference being the addition of language information in the approach to emotion recognition.

The data used is (real-life) spoken human-computer dialogues over the telephone, of which 924 are non-negative and 255 are negative). The approach is based on the ‘information-theoretic notion’ of emotional salience. The task is to recognise negative and non-negative emotions from speech data. The classifiers used are linear discriminant and k-nearest neighbour classifiers.

Emotional keywords are obtained by calculating the emotional salience of the words in the data corpus, where emotional salience is defined as mutual information between words and emotion categories. A set of salient words is selected, and classification is done by maximising the probability of the emotion category given the salient word (multiplying if there are several).

The main results are that adding language information to acoustic improves classification of negative emotion by up to 45.7%, whereas adding acoustic to language information improves classification of negative emotion merely by up to 32.9%.

[Liu et al., 2003] *A Model of Textual Affect Sensing using Real-World Knowledge:*

The data used is the OMCS (Open Mind Common Sense) Corpus in which each fact is represented as one of 20 English sentence patterns. Facts with affective relevance are extracted from the corpus (10% of the total) by keyword spotting (using Ortony et al.’s affective lexicon).

On this basis, a ‘common sense affect model’ is constructed. The model consists of a set of component models which compete with and complement each other. Each model has entries of the form [*frame*][x_1 *happy*, x_2 *sad*, x_3 *anger*, x_4 *fear*, x_5 *disgust*, x_6 *surprise*], $0 < x_i < 1$, where the x_i values represent the salience of the entry to the emotions.

1. subject-verb-object-object model: e.g. [*frame*] = [*<Subject>*: *ep_person_class**, *<Verb>*: *get_into*, *<Object₁>*: *car_accident*, *<Object₂>*:] represents “getting into a car accident can be scary”
2. concept-level unigram model (not a unigram model in the conventional sense): [*<Concept >*: “*car_accident*” [0 *happy*, 0 *sad*, 0 *anger*, ...]
3. concept-level valence model: [*<Concept >*: “*car_accident*”] x , $-1 < x < 1$; but no mapping to emotion categories
4. modifier unigram model: to make up for removing the modifiers in the concept-level unigram model

To construct the models, emotion keywords are propagated in three passes over the corpus. It seems that emotion values initially are 1, then with each propagation are reduced by some factor *d*.

To classify text, it is first segmented into clauses, then ‘linguistically processed’, then evaluated by a 2-stage process using the models. The description of this process is somewhat vague and confusing.

The only evaluation carried out so far involved 20 subjects trying out the EmpathyBuddy email browser in which each sentence is classified according to its predominant emotion, and a cartoon-style face representing that emotion

is displayed next to the sentence. Three settings were tested: one with the emotion classifier described above, one with randomly selected faces, and one with just one, neutral face. The overall user evaluation ranked the first at 5, the second at 4.1, and the third at 3.6 (out of 7), which doesn't reflect too well on how accurately the emotions were recognised.

[Pang et al., 2002] *Thumbs up? Sentiment Classification using Machine Learning Techniques:*

This paper reports results on classification of film reviews into positive or negative. A corpus of 700 negative and 700 positive reviews from 144 different reviewers (available online) was used. A baseline was created by looking at word frequency counts, which yielded an accuracy as high as 69%, with 16% of the remainder undecided.

Three classifiers were tried: Naive Bayes classification, maximum entropy classification, and support vector machines, with 3-fold cross-validation. No stemming or stoplists were used, punctuation was treated as words, negation tags were added to negated words. Each review was represented as a binary vector with 16,165 bits where each bit encodes presence or absence of a word. They don't do any a priori selection of keywords, instead using *all* words that occur. Interestingly, the same word when negated and when not was actually counted as two distinct words.

The best result was 82% accuracy for SVMs and single word features. Using bigrams or POS tags didn't improve results.

[Polzin and Waibel, 2000] *Emotion-sensitive Human-computer Interfaces:*

The data used were 6,000–7,000 speech segments (where a segment is a sentence or utterance with a constant emotion) from movies. Close captions formed the starting point for transcription, transcribers then added missing noises and annotated each segment with three tags: gender, background noise and its predominant emotion. In the experiments 2,991 neutral, 1,586 angry and 1,076 sad segments were used (there weren't enough segments in the other categories). Randomly selected equally sized subsets were used for testing.

The task was to classify into three categories: sad, angry, and neutral. In the human experiments, classification of test data from speech was 70% correct, compared to 55% correct from text.

Automatic classification from speech and text was by selecting the emotion whose model maximises the likelihood of the utterance. Three different kinds of models were used and achieved the following F1-scores: Emotion-specific prosodic models (speech): 60.4%; emotion-specific spectral information model (speech): 63.9%; emotion-specific back-off bigram language models (text): 46.7%.

5 Current Research and Resources

5.1 Research Groups

This section lists some of the more visible project groups with a strong interest in computer-based characters, simulating emotions and in particular how to reflect emotion in NLG. Of course this leaves out many small groups of researchers at numerous institutions that have worked on similar subjects (many of which are mentioned in the following section).

Zoesis Studios: A subset of the Oz Project group took the Oz technology with them and founded Zoesis. They're currently working on extending the technology to emotional language, using Hovy's thesis work.

Headed by E. Bosser, J. Bates, B. Loyall, S. Reilly and P. Weyhrauch.

<http://www.zoesis.com/>

Intelligent User Interfaces at DFKI: headed by W. Wahlster; projects: Presence, SAFIRA, NECA, MagiCster, HUMAINE, MIAU, Verbmobil, SmartKom.

It's hard to group the output of the ongoing research effort on affective computing at **DFKI** into projects — they seem to carry different evolutionary stages of the same software from one project to the next. The **CrossTalk** system (“Meta-Theater with Animated Presentation Agents”) with its emotion engine (also now apparently inserted into the NECA system) is particularly relevant. MIAU platform.

<http://www.dfki.de/iui2/start.html>

Intelligent User Interfaces Group at Bari University: headed by Fiorella de Rosi; projects: GOLEM, XANTHIPPE, HUMAINE, MagiCster.

<http://aos2.di.uniba.it:8080/IntInt.html>

Affective Communication Group (subgroup of Affective Computing) at MIT: headed by Rosalind Picard; research area emotions; focus on emotion recognition; project: Computer Response to User Frustration

http://affect.media.mit.edu/AC_research/communication.html

ITRI, Brighton University: ITRI specialises in NLG, and has two relevant projects: NECA (focus on emotion in computer-based characters), COGENT (will look at emotion-dependent language variation among other dimensions of variation).

<http://www.itri.brighton.ac.uk/>

5.2 Projects

NECA: 2000-2003; project partners: DFKI GmbH (German Research Centre for AI), Freeserve, IPUS (Institute of Phonetics, University of the Saarland), ITRI (Information Technology Research Institute, University of Brighton), OEFAI (Austrian Research Institute for AI, Vienna) (Project Coordinators), Sysis Interactive Simulations AG.

From project homepage: “NECA’s workplan comprises the development of dedicated components, such as components for emotive speech synthesis, generation of combined speech and nonverbal expressions, and components which perform reasoning about a character’s emotional disposition. These components form the building blocks of the generic NECA application platform. This platform will provide the basis upon which specific applications can be built.”

<http://www.ai.univie.ac.at/NECA/>

Oz Project at Carnegie Mellon University: active ca. 1988 to 2002; researchers: Mark Kantrowitz, Joseph Bates, W. Scott Reilly, A. Bryan Loyall, Michael Mateas, Andrew Stern, Peter Weyhrauch, Phoebe Sengers, Sarah Sloane, Matt Glickman.

From Oz homepage: “Oz is a computer system we are developing that allows authors to create and present interactive dramas (Bates 92). [...] The architecture includes a simulated physical world, several characters, an interactor, a theory of presentation, and a drama manager. A model of each character’s body and of the interactor’s body are in the physical world. Outside the physical world, a model of mind controls each character’s actions. The interactor’s actions are controlled by the interactor. Sensory information is passed from the physical world to the interactor through an interface controlled by a theory of presentation. [...] the drama manager influences the characters’ minds, the physical world, and the presentation theory.

Oz has three primary research foci: characters, presentation, and drama. As in traditional media, each of these areas is important for creating a rich dramatic experience. In our research on characters we study how to create computer controlled agents that appear reactive, goal directed, emotional, moderately intelligent, and capable of using natural language (Bates et al. 91, Bates et al. 92, Reilly & Bates 92).”

Despite publications apparently claiming the opposite, selection of content, syntactic structure and lexical items is not connected *at all* to an agent’s emotional state anywhere in the long-running Oz Project.

<http://www-2.cs.cmu.edu/afs/cs.cmu.edu/project/oz/web/oz.html>

SmartKom Project: initial funding 1999-2003; follow-up project to Verbmobil; project consortium: DFKI (main contractor), Daimler Chrysler AG, EML GmbH, University of Erlangen-Nuernberg, ICSI, IMS (Stuttgart University), Munich University, MedialInterface GmbH, Philips GmbH, Siemens AG, Sony GmbH.

From IMS website: “The SmartKom project aims at developing a human-machine interface that is intuitive to use, self-explaining, and adapting to the user’s needs and preferences. The system will recognize speech, gesture and mimic input, and it will generate text, graphics and speech output. The user and the system are intended to use whichever modality is most appropriate for the particular task, the type of information, user preferences, and the application scenario. There are three such scenarios: a home/office working environment, public access to the internet and to information services, and a mobile device (basically a far-advanced cell phone). The project is funded by the German Ministry of Education and Research (BMBF) for the period of September 1999 through August 2003.”

N. Beringer, D. Oppermann, S. Steininger, U. Tuerk, Antje Schweitzer, Norbert Braunschweiler, Edmilson Morais, A. Batliner, Fischer, Wahlster, D. Buehler, W. Minker, J. Haeussler, Sven Krueger, etc.

http://www.smartkom.org/eng/project_en_frames.pl?intro_en.html

GOLEM and XANTHIPPE Projects at University of Bari: Fiorella de Rosis, C. Castelfranchi, R. Falcone, S. Pizzutilo, F. Grasso, I. Poggi.

“The objective of these projects is, on one side, to generate, in life-like agents, personality-rich behaviours; on the other side, to recognise similar behaviours in other agents, such as the user. The way that the behaviour of personality-rich agents is programmed is by defining ‘activation rules’, either in a logical form (Binsted, 1998) or in conditions of uncertainty (Ball and Breese, 1998); these rules define how agents react to a context-driven and/or to an internal emotional or personality state by showing some form of behaviour.”

<http://www.dfki.de/imedia/workshops/i3-spring99/w4-final/roxis.html>

HUMAINE: 2004–2008; EU IST Network of excellence; huge network, 27 partners from 11 countries include: DFKI, Imperial College, King’s College, France Telecom, Oxford University, and Trinity College Dublin.

From Bari page: “The general aim of the network is to equip European teams to develop systems that register, model and/or influence human emotional and emotion-related states. We call these emotion-oriented systems. The network will underpin sound long-term development of these systems by clarifying the scientific, cultural and ethical underpinnings of emotion-oriented computing; developing accessible literature and standards based on that understanding; and establishing enduring relationships among research groups in the key areas of information technology and the science of emotion.”

<http://emotion-research.net/>

MagiCster: 2000–2003; EC IST Project; project partners: University of Edinburgh, Division of Informatics, Università degli Studi di Roma “La Sapienza”, Dipartimento di Informatica e Sistemistica, Deutsches Forschungszentrum für Künstliche Intelligenz, Intelligent User Interfaces Department, Swedish Institute of Computer Science, Università degli Studi di Bari, Research Group on Intelligent Interfaces, AvatarMe.

Objectives (from Bari project page): “(i) to design a believable conversational interface agent which makes use of gaze, facial expression, gesture and body posture as well as speech in a synchronised fashion; (ii) to evaluate the use of the conversational agent in laboratory conditions to determine which aspects of the embodied agent are important for what types of human-computer interaction; (iii) 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.”

<http://www.ltg.ed.ac.uk/magicster/>

Presence Project at DFKI: (homepage last updated 2000).

From homepage: “Presence is an in-house initiative timed to catch the wave of recent academic and commercial interest in life-like characters. The primary goal of the project is to advance our understanding of the fundamental technology needed to drive Life-Like characters by combining the skills and expertise present in the Multiagent Systems and Intelligent User Interface groups.

The Presence project will use lifelike characters as virtual receptionists / infotainers / accompanying guides for visitors to the German Research Centre for Artificial Intelligence (DFKI GmbH). Here we will explore the hypothesis that using an explicit affective model (of both agent and user) to guide the presentation strategies used in the human-agent conversational dialogue will (a) create a more natural and intuitive user interface (by tailoring the conversation to an individual person); (b) provide the user with an engaging and enjoyable experience; and (c) enhance the believability of virtual characters.”

Steve Allen, Elisabeth Andr, Patrick Gebhard, Wenji Mao.

<http://www.dfki.de/allen/Presence.html>

Mission Rehearsal Exercise virtual training environment project: homepage lists 42 project members, NLG team mainly Traum, Fleischman and Hovy); see notes on papers above.

http://www.ict.usc.edu/disp.php?bd=proj_mre

SAFIRA: Supporting Affective Interactions for Real-time Applications. 2000–2003. EU IST Project. Partners: Intelligent Agents & Synthetic Characters Group (Lisbon), ADETTI (Lisbon), DFKI, Fraunhofer IMK, Imperial College (London), FAI, Swedish Institute of Computer Science.

From project homepage: “The project objective is to bring to the software community an enabling technology to support affective interactions, in particular:

- * To create a framework to enrich interactions and applications with an affective dimension;
- * To implement a toolkit for affective computing combining a set of components addressing affective knowledge acquisition, representation, reasoning, planning, communication and expression;
- * 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.”

<http://gaips.inesc.pt/safira>

Emotion in Speech Project: joint research project between the Speech Laboratory at the University of Reading and the Department of Psychology at the University of Leeds, and was funded by the Ministry of Defence (DERA, Malvern) and ESRC.

P. Roach, R. Stibbard, S. Arnfield, J. Osborne, J. Setter (Reading). P. Greasley, M. Waterman, C. Sherrard (Leeds).

<http://www.rdg.ac.uk/AcaDepts/11/speechlab/emotion/>

JST/CREST-ESP Project: ATR Human Information Science Laboratories in the Keihanna Science City, Kyoto, Japan. JST ESP Project (Japan Science and Technology Agency Expressive Speech Processing):

“The purpose of this research is to model the way that speakers use different forms of intonation and voice quality to express levels of meaning, above and beyond that which is expressed by the words of an utterance alone. The research will determine optimal features that can be reliably extracted from a speech waveform, and map them onto

known linguistic structures and identifiable speech-act characteristics so that the desired intention of the speaker can be expressed or interpreted. The mapping will be two-way.”

Nick Campbell, Kiyohiro Shikano, Hideki Kashioka.

ERMIS Project: (Emotionally Rich Man-Machine Interaction Systems). Jan 2002–Dec 2005. EU IST Project. 14 partners including British Telecom, France Telecom, Queen’s University Belfast, King’s College London.

“The main objective of the ERMIS project is the development of a prototype system for human computer interaction that can interpret its users’ attitude or emotional state, e.g., activation/interest, boredom, and anger, in terms of their speech and/or their facial gestures and expressions. The adopted technologies include linguistic and paralinguistic speech analysis and robust speech recognition, facial expression analysis, interpretation of the user’s emotional state using hybrid, neurofuzzy, techniques, while being in accordance with the MPEG-4 standard. Specific attention is given to the evaluation of the system’s ability to improve effectiveness, user friendliness and user satisfaction, while examining and resolving related ethical issues.”

Scope: The development of a prototype system for human computer interaction than can interpret its users’ attitude or emotional state, e.g., activation/interest, boredom, and anger, in terms of their speech and/or their facial gestures and expressions

Adopted technologies: Linguistic and paralinguistic speech analysis and robust speech recognition, facial expression analysis, interpretation of the user’s emotional state using hybrid, neurofuzzy, techniques, while being in accordance with the MPEG-4 standard.

Stefanos Kollias, Luc van Gool.

<http://www.image.ntua.gr/ermis/>

ITR Project: University of Illinois; “The goal of this project is to contribute to the development of a human-computer interaction environment in which the computer detects and tracks the user’s emotional, motivational, cognitive and task states, and initiates communications based on this knowledge, rather than simply responding to user commands.”

<http://itr.beckman.uiuc.edu/index.html>

Meeting Recorder Project: Partners: ICSI speech group at Berkeley, SSLI lab at University of Washington, SRI’s STAR Lab. Funding from DARPA, and IBM.

From webpage: “Despite recent advances in speech recognition technology, successful recognition is limited to co-operative speakers using close-talking microphones. There are, however, many other situations in which speech recognition would be useful - for instance to provide transcripts of meetings or other archive audio. Speech researchers at ICSI, UW, SRI, and IBM are very interested in new application domains of this kind, and we have begun to work with recorded meeting data. [...]A key issue in the project is to specify the goals and applications. While the basic idea is to develop recognition that could transcribe conventional meetings, this would be useful only in so far as it would support applications such as searching for particular information or producing automatic summaries. “

<http://www.icsi.berkeley.edu/Speech/mr/mtgrcdr.html>

PHYSTA: 1998–2001. EU TMR Project. King’s College London, Nijmegen University, Milan University, Queen’s University Belfast, Image, Video and Multimedia Systems Laboratory.

<http://www.image.ece.ntua.gr/physta/>

Others: Virtual Human (Germany); PF-STAR (EU); SIMILAR Network of Excellence (EU).

5.3 Corpora and Other Data Collections

Lee et al. at University of Southern California: corpus of (real-life) spoken human-computer dialogue over telephone; each utterance is labelled negative/non-negative and male/female; there are 924 non-negative and 255 negative utterances.

Amities Project at Sheffield University (<http://www.dcs.shef.ac.uk/research/groups/nlp/amities/>): corpus of 100 human (agent-client) dialogues recorded in a stock exchange service center; annotated with (i) DAMSL-like dialogue labels (Allen et al.) three levels of labelling for dialogue labels: information, forward-looking, backward-looking; (ii) dialogue progression axis labels (Devillers et al.); represented on two axes: progression and accidents; and (iii) emotion labels: anger, fear, satisfaction, excuse, neutral.

Emotion in Speech Corpus produced in Emotion in Speech Project at Reading and Leeds Universities (not distributed). Annotation in ToBI [Silverman et al., 1992], with additional annotation in the “miscellaneous” tier of the formalism.

Belfast Naturalistic Emotion Database: database of emotional speech, containing video clips, .wav files for main speaker for each clip, and orthographic transcriptions of .wav files. The transcriptions are made available under certain circumstances.

ISLE Natural Interactivity and Multimodality Working Group Deliverable D8.1: Survey of NIMM Data Resources, Current and Future User Profiles, Markets and User Needs for NIMM Resources: division into facial expression resources and gesture resources - there may be some transcribed speech in some of the resources

SmartKom Project Corpora: there is one small corpus annotated for speech, gestures, and emotions collected Wizard of Oz style; report authors: Beringer, Oppermann, Steininger at Munich University

Also Fischer and Batliner et al. mention the aim of recording 70 human-computer dialogues each 5 mins in length, annotated for lexical, conversational and prosodic peculiarities; this is a unique and potentially very useful approach, because it doesn't use emotion labels (which are very subjective and usually forced-choice) but annotate properties of the language combinations of which are assumed to correlate with emotions.

ITR Project Corpus: Dan Roth et al. collecting corpus of video/audio/text annotated for emotion. In early 2003 had collected data from about 40 children carrying out simple tasks. Were planning to make corpus available.

Columbia Emotional Speech Corpus: NLP group at Columbia, corpus is annotated by volunteers online:
<http://www1.cs.columbia.edu/nlp/emotion/>

DARPA Communicator project corpus: natural human-computer dialogue over the telephone, making travel arrangements (or rather, pretending to).

TMR PHYSTA Project: the report on existing resources lists only 3 speech databases with emotion annotation (and no others), two of which is material read by actors, and the other is the Emotion in Speech Corpus above.

Corpus of film reviews used in Pang et al. (2002) as summarised above.

Other possible sources of emotion data: text and stage directions in drama; football commentaries with rising levels of excitement (useful for excitement yes/no vs. intensity of excitement); transcribed football commentaries: e.g. Mirjam Wester et al. in MUMIS project; TV and video subtitles; newsgroups debates (and similar online resources).

6 Conclusions

The interest in what has become known as 'affective computing' has exploded since 1995, although even Picard's exercise in ground preparation [Picard, 1995] was already able to cite a survey paper on emotional speech synthesis [Murray and Arnott, 1993], and a survey of AI models of emotion [Pfeifer, 1988]. Research on emotional variation in NLG goes back at least as far as Klein et al.'s work on the Automatic Novel Writer in the early 70s, but very little has been achieved in almost two decades since then. In particular, the body of research on how to express emotion in the automatically generated word string has remained very small (Section 2).

Yet there is plenty of evidence that the word string does play an important role in human and machine detection of emotion (Section 3). Emotionally varied NLG must therefore play a role in any research that aims to make human-computer interaction more 'emotionally intelligent'.

The problem has not been that researchers haven't aimed to make emotionally varied NLG possible, as many project outlines testify (Sections 4 and 5), but that the results have tended to be small-scale, specialised and often fragmented techniques rather than comprehensive methodologies. Perhaps the biggest obstacle is the lack of a strong basis in the form of a substantial body of research on emotion as expressed in the word string, in either psychological or linguistics research.

Automatic emotion identification methods provide a handle on how to make up for this lack (Section 3). They can be used to identify lexical, syntactic and other properties that are correlated with and help to identify emotion. They can also provide the empirical grounding that is at the moment completely absent in emotional language generation.

Given the strong interest in AI in 'believable agents' and 'affective computing' that has grown exponentially over at least the last decade, it comes as a surprise to discover just how little has been achieved (or even attempted) in emotional NLG. Apart from a single notable exception (Hovy's PAULINE system) the field is wide open: we're not anywhere close to creating NLG systems capable of systematic emotional variation.

7 Acknowledgements

I am grateful to all those who responded to my queries on the SIGGen and Corpora mailing lists, in particular to Ellen Douglas-Cowie and Edelle McMahon for letting me look at transcriptions of the Belfast Naturalistic Emotion Database, Sheldon Klein for alerting me to the Automatic Novel Writer, David McDonald for comments on the state of emotional NLG, Fiorella de Rosis for additional explanations of how the MagiCster system works, and Dan Roth for pointing me to the ITR corpus. Of course, any mistakes in this report are solely my own responsibility.

References

- [André et al., 2000] André, E., Rist, T., van Mulken, S., Klesen, M., and Baldes, S. (2000). The automated design of believable dialogues for animated presentation teams. In J. Cassell, J. Sullivan, S. P. and Churchill, E., editors, *Embodied Conversational Agents*, pages 220–255. MIT Press.
- [Ang et al., 2002] Ang, J., Dhillon, R., Krupski, A., Shriberg, E., and Stolcke, A. (2002). Prosody-based automatic detection of annoyance and frustration in human-computer dialog. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP 2002)*, volume 3, pages 2037–2040.
- [Bartlett et al., 1999] Bartlett, M. S., Hager, J. C., Ekman, P., and Sejnowski, T. J. (1999). Measuring facial expressions by computer image analysis. *Psychophysiology*, 36:253–263.
- [Batliner et al., 2000] Batliner, A., Fischer, K., Huber, R., Spilker, J., and Nöth, E. (2000). Desperately seeking emotions: Actors, wizards, and human beings. In *Proceedings of the ISCA workshop on Speech and Emotion: A Conceptual Framework for Research*, pages 195–200.
- [Breese and Ball, 1998] Breese, J. and Ball, G. (1998). Modeling emotional state and personality for conversational agents. Technical Report MSR-TR-98-41, Microsoft Research, Advanced Technology Division.
- [Cahn, 1990] Cahn, J. (1990). The generation of affect in synthesized speech. *Journal of the American Voice I/O Society*, 8:1–19.
- [Carolis et al., 2002] Carolis, B. D., Carofiglio, V., and Pelachaud, C. (2002). From discourse plans to believable behavior generation. In *Proceedings of the 2nd International Conference on Natural Language Generation*, pages 65–72.
- [de Rosis and Grasso, 2000] de Rosis, F. and Grasso, F. (2000). Affective natural language generation. In Paiva, A. M., editor, *Affective Interactions*, number 1814 in Springer Lecture Notes in AI, pages 204–218. Springer.
- [Devillers et al., 2002] Devillers, L., Vasilescu, I., and Lamel, L. (2002). Annotation and detection of emotion in a task-oriented human-human dialog corpus. In *Proceedings of ISLE Workshop*.
- [Dryer, 1999] Dryer, D. C. (1999). Getting personal with computers: How to design personalities for agents. *Applied Artificial Intelligence*, 13(3):273–295.
- [Elliot, 1992] Elliot, C. (1992). *The Affective Reasoner: A Process Model of Emotions in a Multi-agent System*. PhD thesis, Institute for the Learning Sciences, Northwestern University. Available as Technical Report No. 32.
- [Fleischman and Hovy, 2002] Fleischman, M. and Hovy, E. (2002). Towards emotional variation in speech-based natural language generation. In *Proceedings of the Second International Natural Language Generation Conference (INLG02)*, pages 57–64.
- [Goertzel et al., 2000] Goertzel, B., Silverman, K., Hartley, C., Bugaj, S., and Ross, M. (2000). The baby webmind project. In *Proceedings of AISB 2000*.
- [Gratch, 2000] Gratch, J. (2000). Modeling the interplay between emotion and decision making. In *Proceedings of the 9th Conference on Computer Generated Forces and Behavioral Representation*.
- [Haimowitz, 1991] Haimowitz, I. (1991). Modeling all dialogue system participants to generate empathetic responses. *Computer Methods and Programs in Biomedicine*, 35:321–330.
- [Healey, 2000] Healey, J. (2000). *Wearable and Automotive Systems for Affect Recognition from Physiology*. PhD thesis, Mass. Inst. Technology. Technical Report 526.
- [Holzman and Pottenger, 2003] Holzman, L. and Pottenger, W. (2003). Classification of emotions in internet chat. Technical Report LU-CSE-03-002, Dept of Science and Engineering, Lehigh University.
- [Hovy, 1986] Hovy, E. (1986). Putting affect into text. In *Proceedings of the 8th Conference of the Cognitive Science Society*.
- [Hovy, 1988] Hovy, E. H. (1988). *Generating Natural Language Under Pragmatic Constraints*. Lawrence Erlbaum, Hillsdale, New Jersey.
- [Hovy, 1990] Hovy, E. H. (1990). Pragmatics and natural language generation. *Artificial Intelligence*, 43:153–197.

- [Klein et al., 1973] Klein, S., Aeschlimann, J., Balsiger, D. F., Converse, S. L., Court, C., Foster, M., Lao, R., Oakely, J. D., and Smith, J. D. (1973). Automatic novel writing. Technical Report UWCS Tech Report No. 186, Department of Computer Sciences, University of Wisconsin Madison. An abridged version also appears in, *Text Processing/Textverarbeitung*. Edited by W. Burghardt & K. Hlker, pp. 338–412, Berlin & New York: Walter de Gruyter, 1979.
- [Kölln, 1995] Kölln, M. E. (1995). Employing user attitudes in text planning. In *Proceedings of the 5th European Workshop on Natural Language Generation*, pages 163–179.
- [Landauer and Dumais, 1997] Landauer, T. K. and Dumais, S. T. (1997). A solution to plato’s problem: The latent semantic analysis theory of the acquisition, induction and representation of knowledge. *Psychological Review*, 104:211–140.
- [Lee et al., 2001] Lee, C. M., Narayanan, S., and Pieraccini, R. (2001). Recognition of negative emotions from the speech signal. In *Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop (ASRU 2001)*, pages ??–??
- [Lee et al., 2002] Lee, C. M., Narayanan, S., and Pieraccini, R. (2002). Combining acoustic and language information for emotion recognition. In *Proc. ICSLP’02*, pages ??–??
- [Li and Zhao, 1998] Li, Y. and Zhao, Y. (1998). Recognizing emotions in speech using short-term and long-term features. In *Proceedings of ICSLP 1998*, pages 2007–2010.
- [Liu et al., 2003] Liu, H., Lieberman, H., and Selker, T. (2003). A model of textual affect sensing using real-world knowledge. In *Proceedings of the Seventh International Conference on Intelligent User Interfaces (IUI 2003)*, pages 125–132.
- [Loyall and Bates, 1997] Loyall, A. B. and Bates, J. (1997). Personality-rich believable agents that use language. In *Proceedings of the first International Conference on Autonomous Agents*.
- [Moon and Nass, 1996] Moon, Y. and Nass, C. (1996). How ”real” are computer personalities? *Communication Research*, 23(6):651–674.
- [Murray and Arnott, 1993] Murray, I. R. and Arnott, J. L. (1993). Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion. *Journal of the Acoustic Society of America*, 93:1097–1108.
- [Nass et al., 1994] Nass, C., Steuer, J., and Tauber, E. R. (1994). Computers are social actors. In *Proceeding of the CHI Conference*.
- [Nass and Sundar, 1994] Nass, C. I. and Sundar, S. S. (1994). Is human-computer interaction social or parasocial? *Submitted to Human Computer Interaction*.
- [Ortony et al., 1988] Ortony, A., Clore, G., and Collins, A. (1988). *The Cognitive Structure of Emotions*. Cambridge University Press.
- [Pang et al., 2002] Pang, B., Lee, L., and Vaithyanathan, S. (2002). Thumbs up? sentiment classification using machine learning techniques. In *Proceedings of the 2000 Conference on Empirical Methods in Natural Language Processing*.
- [Pfeifer, 1988] Pfeifer, R. (1988). Artificial intelligence models of emotion. In Hamilton, V., Bower, G. H., and Frijda, N. H., editors, *Cognitive Perspectives on Emotion and Motivation*, volume 44 of *Behavioural and Social Sciences*, pages 287–320. Kluwer.
- [Picard, 1995] Picard, R. (1995). Affective computing. Technical Report Perceptual Computing TR 321, MIT Media Lab.
- [Picard, 1997] Picard, R. W. (1997). *Affective Computing*. MIT Press, Cambridge, Massachusetts.
- [Polzin and Waibel, 2000] Polzin, T. and Waibel, A. (2000). Emotion-sensitive human-computer interfaces. In *Proceedings of the ISCA-Workshop on Speech and Emotion*.
- [Prendinger and Ishizuka, 2001] Prendinger, H. and Ishizuka, M. (2001). Agents that talk back (sometimes): Filter programs for affective communication. In *Second Workshop on Attitude, Personality and Emotions in User-adapted Interaction*.

- [Reeves and Nass, 1996] Reeves, B. and Nass, C. (1996). *The Media Equation*. CUP.
- [Schröder, 2001] Schröder, M. (2001). Emotional speech synthesis — a review. In *Proceedings of Eurospeech 2001*, volume 1, pages 561–564.
- [Silverman et al., 1992] Silverman, K., Beckman, M., Pitrelli, J., Ostendorf, M., Wightman, C., Price, P., Pierrehumbert, J., and Hirschberg, J. (1992). ToBI: A standard for labelling english prosody. In *Proceedings of ICSLP 92*, volume 2, pages 867–870.
- [Vyzas and Picard, 1999] Vyzas, E. and Picard, R. (1999). Offline and online recognition of emotion expression from physiological data. In *Proc. Workshop Emotion-Based Agent Architectures, Third Int'l Conf. Autonomous Agents*, pages 135–142.
- [Walker, 1992] Walker, M. (1992). Redundancy in collaborative dialogue. In *Proceedings of the 15th International Conference on Computational Linguistics (COLING '92)*.
- [Walker, 1996] Walker, M. (1996). The effect of resource limits and task complexity on collaborative planning in dialogue. *Artificial Intelligence*, 85(1-2):181–243.
- [Walker et al., 1997] Walker, M. A., Cahn, J. E., and Whittaker, S. J. (1997). Improvising linguistic style: Social and affective bases of agent personality. In *Proceedings of the First International Conference on Autonomous Agents*, pages 96–105.
- [Whiteside, 1998] Whiteside, S. P. (1998). Simulated emotions: An acoustic study of voice and perturbation measures. In *Proceedings of ICSLP 1998*, pages 699–703.
- [Zhou et al., 1998] Zhou, G., Hansen, J. H. L., and Kaiser, J. F. (1998). Linear and nonlinear speech feature analysis for stress classification. In *Proceedings of ICSLP 1998*.